



universität  
wien

# **DIPLOMARBEIT**

**Titel der Diplomarbeit:**

**„Illustrative Testkonstruktionskritik des HAWIK-IV  
am Beispiel der Hochbegabungsdiagnostik“**

**Verfasser:**

**Georg Wilflinger**

**angestrebter akademischer Grad:**

**Magister der Naturwissenschaften (Mag.rer.nat.)**

**Wien, im Juni 2009**

**Studienkennzahl lt. Studienblatt: A298**

**Studienrichtung lt. Studienblatt: Psychologie**

**Betreuer: Univ. Prof. Dr. Mag. Klaus Kubinger**



**MEINEN GROSSARTIGEN TÖCHTERN  
LEONIE, HANNAH UND MINOU GEWIDMET.**



## **VORWORT**

Mein Dank gilt Univ. Prof. Klaus Kubinger, MMag. Silvia Schubhart und Dr. Stefana Holocher-Ertl, die viel zum Entstehen dieser Arbeit beigetragen haben: für ihre inhaltlichen Anregungen, Aufmunterungen, ihre Anerkennung, Kritik und ihre Hilfe.

Weiters danke ich meinen Kolleginnen Anna und Grete, die mich mit Rat und Freundschaft durch das Psychologiestudium begleitet und so ganz wesentlich zum Gelingen beigetragen haben, für das Korrekturlesen.

Meinen Eltern möchte ich an dieser Stelle danken, dass sie mir durch ihre große Unterstützung das Studium ermöglicht haben, und Hannah, Manuela, Leonie, Johanna, Minou und Eva für ihr Verständnis und die Zeit, die ich fürs Studium und diese Arbeit brauchte.



**Abstract:** In dieser Arbeit wird die Eignung der Intelligenztestbatterie HAWIK-IV für die Einzelfalldiagnostik am Beispiel der Hochbegabungsdiagnostik anhand einer anfallenden Stichprobe von 41 Kindern und Jugendlichen im Alter zwischen 7;10 und 16;4 Jahren untersucht und diskutiert. Dabei werden sowohl testtheoretische Aspekte (wie Validität, Reliabilität und Fairness), als auch die Eignung im Sinne der praktischen Handhabung untersucht. Die statistischen Untersuchungen sind als explorative Datenauswertung zu verstehen und stellen keine Hypothesenprüfungen dar. Die Ergebnisse zeigen hinsichtlich der Konstruktvalidität, dass die faktorenanalytisch begründete Zuordnung der Untertests zu Indizes anhand der vorliegenden Daten nur teilweise bestätigt werden kann, und in einigen Fällen weder inhaltlich noch aus den Daten nachvollziehbar erscheint. Letztere Kritik betrifft den Index *Wahrnehmungsgebundenes Logisches Denken* und den Untertest *Rechnerisches Denken*. Darüber hinaus zeigen sich Hinweise für eine Konfundierung der dreikategoriell verrechneten Untertests des Index *Sprachverständnis* mit Persönlichkeitsfaktoren und eine ungenaue Skalierung dieser Untertests. Des weiteren erscheinen einige Untertests als zu wenig messgenau, um als Einzelskala im Rahmen eines Profils interpretiert zu werden. Dies zeigt sich sowohl in den Werten der kritisierten Split-half-Reliabilitätskoeffizienten wie auch in den Itemcharakteristika Trennschärfe und Itemschwierigkeit. Illustrativ werden Verstöße gegen das Kriterium der Fairness aufgrund der Abbruchregeln aufgezeigt, die sich teilweise durch die Datenanalyse belegen lassen. Aus diesen Gründen wird der HAWIK-IV für eine differenzierte Intelligenzprofil Diagnostik, wie es beispielsweise das „Wiener Diagnosemodells zum Hochleistungspotenzial“ fordert, als nicht empfehlenswert bewertet.

**Abstract:** This thesis evaluates and discusses the suitability of the German WISC-IV for identifying intellectually gifted children, using a sample of 41 children between the age of 7;10 and 16;4. Test-theoretical aspects were examined (concerning test validity, reliability and fairness) as well as the suitability of the test in terms of practicability in handling. Although no hypothesis tests were done, statistics were calculated as an explorative data analysis. The results regarding construct validity show that the factor-analytically based allocation of the subtests to indices can only be partly confirmed. This criticism is valid for the *Perceptual Reasoning Index* and the subtest *Arithmetic*. Beyond that indications are shown, that the scalings of the three subtests of the *Verbal*

*Comprehension Index* are not sufficient as well as confounded with personality factors. Some other subtests seem unsuitable for individual interpretation as part of a profile due to their weak reliability. The same is valid for split-half-reliability, item-characteristics, and for the corrected item-total correlation or difficulty. Possible violations of fairness due to the discontinuation rules of the test are illustrated and can be partly confirmed by data analysis. For the discussed reasons the German WISC-IV is evaluated as not recommendable for differentiated intelligence diagnostics, as for example the “Viennese model for assessment of high achievement potential“ requires.



# INHALT

<b>I</b>	<b>EINLEITUNG .....</b>	<b>11</b>
<b>II</b>	<b>THEORETISCHER TEIL .....</b>	<b>15</b>
1	DIAGNOSTIK INTELEKTUELLER HOCHBEGABUNG.....	15
1.1	Anforderungen des „Wiener Diagnosemodells zum Hochleistungspotenzial“ an ein Verfahren zur Intelligenzmessung .....	16
<b>III</b>	<b>EMPIRISCHER TEIL.....</b>	<b>19</b>
2	ZIELSETZUNG UND BESONDERHEITEN DER EMPIRISCHEN UNTERSUCHUNG.....	19
3	VALIDITÄT.....	23
3.1	Intelligenztheoretische Fundierung des HAWIK-IV .....	24
3.2	Konstruktvalidität.....	25
3.3	Eigene Untersuchungen zur Validität (Faktorenanalysen) .....	29
3.4	Diskussion und Zusammenfassung der Ergebnisse zur Validität .....	34
4	KRITIK AN DER DREIKATEGORIELLEN VERRECHNUNG DER UNTERTESTS GF, WT UND AV .....	38
4.2	Inhaltliche Kritikpunkte (hinsichtlich Validität).....	39
4.3	Statistische Untersuchungen: „2-Punkt-Antworten als eigene Dimension“ .....	43
4.4	Zusammenhänge: 2-Punkt-Antworten und Persönlichkeitsvariablen.....	45
4.5	Diskussion und Zusammenfassung: Kritik hinsichtlich Validität.....	50
4.6	Kritik an der dreikategoriellen Verrechnung hinsichtlich Skalierung .....	51
4.7	Kritik hinsichtlich der praktischen Handhabung .....	55
5	RELIABILITÄT .....	56
5.1	Split-half-Reliabilität.....	56
5.2	Trennschärfe und Itemschwierigkeit.....	60
5.3	<i>Mosaik-Test</i> - Reliabilität, Itemtrennschärfe, Itemschwierigkeit.....	62
5.4	<i>Bildkonzepte</i> - Reliabilität, Trennschärfe, Itemschwierigkeit.....	67
5.5	<i>Matrizen-Test</i> - Reliabilität, Trennschärfe, Itemschwierigkeit .....	70

5.6	<i>Gemeinsamkeiten Finden</i> – Reliabilität, Trennschärfe, Itemschwierigkeit.....	73
5.7	<i>Wortschatz-Test</i> – Reliabilität, Trennschärfe, Itemschwierigkeit .....	76
5.8	<i>Allgemeines Verständnis</i> – Reliabilität, Trennschärfe, Itemschwierigkeit.....	81
5.9	<i>Rechnerisches Denken</i> – Reliabilität, Trennschärfe, Itemschwierigkeit .....	83
5.10	<i>Allgemeines Wissen</i> – Reliabilität, Trennschärfe, Itemschwierigkeit.....	86
5.11	Reliabilität der Untertests ZST, ZN, BZF und SYS .....	88
5.12	Empirische Split-half-Reliabilitäten der dichotomisierten UTs GF, AV und WT ....	88
5.13	Diskussion und Zusammenfassung der Ergebnisse zur Reliabilität.....	97
6	FAIRNESS.....	99
6.1	Die Abbruchregeln .....	100
6.2	Zusammenhänge der durchgeführten Abbrüche mit Persönlichkeitseigenschaften.....	102
6.3	Relative Lösungshäufigkeiten im Sinne des Abbruchkriteriums (je Untertest).....	108
6.4	Darstellung der durchgeführten Abbrüche hinsichtlich Fairness (je Untertest).....	114
6.5	Zusammenfassung: Abbruchregeln und Fairness.....	134
7	DISKUSSION.....	136
8	ZUSAMMENFASSUNG .....	139
9	LITERATUR .....	143
10	ANHANG .....	144
	Lebenslauf .....	147

## I EINLEITUNG

Ich erlaube mir, mit einem Cartoon zu beginnen: Calvin, die Comicfigur, ist ein Bub, der mit Einfallsreichtum und Fantasie und jeder Menge Eigensinn die Erwachsenen seiner Umgebung stark fordert, und trotz seiner Schläue (fast) zum Schulversager wird.



(Watterson; 1993, in: Grosjean, 2003, S. 69; grafisch nachbearbeitet)

Grosjean (2003), dessen kritischer Auseinandersetzung mit der akademischen Psychologie aus der Sicht eines Studenten dieser Comicstrip entnommen wurde, schreibt dazu:

„In dem Cartoon sieht man sehr genau, warum Calvin von seinen „Lehrern“ – sei es zu Hause oder in der Schule – als „schlechter Schüler“ eingestuft wird: Das Problem ist überhaupt nicht, dass seine Antworten irgendwie einen Mangel an „Intelligenz“ aufzeigen (man muss im Gegenteil sagen, dass sie für einen Sechsjährigen ziemlich listig sind); aber es sind leider nicht die Antworten, welche die „Lehrer“ hören wollen.“ (Grosjean, 2003, S. 69)

Ohne auf die schier endlose Auseinandersetzung über Definitionen und Modelle von Intelligenz genauer einzugehen, ist festzustellen, dass es keine verbindlichen Definitionen oder Ansichten darüber gibt, was Intelligenz genau ist oder umfasst; Intelligenz es ist eben ein psychologisches *Konstrukt*. Dies ist eigentlich dahingehend zu erweitern, dass es sich dabei nicht um ein, sondern eher um relative *viele* Konstrukte handelt, die alle den gleichen Namen beanspruchen und zumindest ähnliches beinhalten. Nach Boring (1923, z.B. in Amelang, Bartussek, Stemmler & Hagemann, 2006) ist Intelligenz eben das, was ein Intelligenztest misst; in der Ironie dieser Aussage wird durchaus auch auf die Beliebigkeit des Gebrauchs dieses Begriffes angespielt, und gleichzeitig werden damit sehr weitreichende Problembereiche umrissen: Die Höhe der

„Intelligenz“ als Eigenschaft, die einer Person<sup>1</sup> zugeschrieben wird, ist in erster Linie durch den Betrachter, bzw., wenn es als Ergebnis eines psychologisch-diagnostischen Verfahrens verstanden wird, durch den Test definiert, mit dem eben diese „Intelligenz“ festgestellt wurde; andererseits ist die Zuschreibung von Intelligenz auch ein (soziales, wie auch praktisch-psychologisches) Faktum, das den Eindruck erweckt, es gäbe eben diese *eine* (und auch als solche feststellbare!) Intelligenz, die Personen in je unterschiedlichem Ausmaß besäßen, woraus oftmals Entscheidungen (hinsichtlich Schulwahl, Bewerberauswahl...) abgeleitet werden.

Die notwendige Differenzierung der festgestellten Intelligenz anhand ihrer Operationalisierung, d.h. anhand ihres Messinstruments, geschieht denn in der Praxis nur allzu selten: sowohl in Fragen der Beschulung wie auch der Mitarbeiterauswahl, in der Beratung, ja auch in der klinischen Klassifikation und Diagnostik werden Entscheidungen unter Verwendung der festgestellten Höhe der Intelligenz getroffen, so als wäre klar definiert, was das genau sei.<sup>2</sup> Wechslers Definition (1944, zitiert nach Amelang et al, 2006, S. 166), dass „Intelligenz [...] die zusammengesetzte oder globale Fähigkeit [...] sei], zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinander zu setzen“, erreichte eine große Verbreitung und scheint umfassend und gleichzeitig unbestimmt genug, um für die meisten Verwendungen des Begriffs „Intelligenz“ zu passen. Doch welcher Test könnte dieses umfassende Konstrukt messen? Man denke nur an die Vielfalt möglicher Umgebungen und an die noch größere Zahl vernünftiger und zweckvoller Auseinandersetzungen mit diesen! Im Sinne dieser Definition wurden viele Intelligenztests entwickelt, zu denen auch der im Rahmen dieser Arbeit diskutierte HAWIK-IV (*Hamburg-Wechsler-Intelligenztest für Kinder-IV. Übersetzung und Adaption der WISC-IV® von David Wechsler*. Petermann & Petermann, 2007) gehört.

Doch trotz des Vorliegens von Intelligenztests, die eine konkrete Umsetzung der Vorgaben dieser als geeignet anerkannten Beschreibung bzw. Definition darzustellen

---

<sup>1</sup> Für Personengruppen, die in der vorliegenden Arbeit mit geschlechtssensitiven Begriffen bezeichnet werden, wird immer die grammatikalisch männliche Form verwendet, insofern beide Geschlechter damit gemeint sind; weibliche Repräsentanten dieser Personengruppen – beispielsweise Psychologinnen – sind dabei miteingeschlossen.

<sup>2</sup> So fordert das ICD-10 beispielsweise in Hinblick auf die Art der Intelligenzmessungen (in Fragen der Intelligenzminderung) nur: „Der IQ sollte anhand von standardisierten, auf die jeweiligen kulturellen Gegebenheiten adaptierten, individuell angewandten Intelligenztests bestimmt werden.“ ohne zu beschreiben, welche intelligenzmäßigen Fertigkeiten dabei zu erfassen wären. (Dilling, Mombour & Schmidt, 2000, S. 255)

scheinen, stellen sich für den einzelnen Psychologen in Hinblick auf seine praktische diagnostische Tätigkeit wichtige Fragen, beispielsweise sind dies folgende:

Wie kann ich Intelligenz im Sinne *einer* Fähigkeit verantwortungsvoll „testen“ und dieses Ergebnis („IQ“) weitergeben oder interpretieren, obwohl ich doch weiß, dass kein Test das ganze Spektrum dieses umfassenden (und sehr wertenden!) Konstruktes erfassen kann?

Was genau ist es denn, was im Rahmen der von mir verwendeten Intelligenztests erfasst wird? Was *genau* ist es denn? Und: Erfasst er es auch (mess-) *genau genug*?

Und letztlich kumuliert es in der Frage: Wird der Test denn der Intelligenz dieser Testperson gerecht? Benachteiligt er sie nicht etwa, prüft er denn tatsächlich die Fähigkeiten, die für die vernünftige Auseinandersetzung eben dieser Person mit eben ihrer individuellen Umgebung notwendig sind?

Schon aus der Formulierung dieser Fragen wird klar, dass es sich dabei nicht um rein „akademische“ Fragen handelt. Es sind sehr lebensrelevante Themen mit zum Teil weitreichenden Folgen: sowohl für die getesteten Personen als auch, und das darf nicht außer Acht gelassen werden, für die testenden Psychologen und letztendlich auch für die Psychologie als Profession. Und obwohl diese Fragen alles andere als „rein akademisch“ sind, ist es möglich und notwendig, ihnen auch mit akademischen Mitteln nachzugehen. Im konkreten Fall sind dies auch Mittel der Statistik: dies geschieht nicht in der Annahme des Verfassers, diese Fragen ließen sich mit Mitteln der Statistik bestätigend beantworten; es geschieht im Bestreben, die möglicherweise fehlende psychologische Güte auch hinter dem vorteilhaften Bild der publizierten testtheoretischen *Gütekriterien* aufzuzeigen; dies macht es aber notwendig, eben auch die testtheoretische Fundierung kritisch zu diskutieren. Die Fragen, denen diese Arbeit demnach nachgeht, sind die nach der Güte des HAWIK-IV, vor allem: nach seiner Validität, seiner Reliabilität und seiner Fairness.<sup>3</sup>

---

<sup>3</sup> Die erste der oben gestellten Fragen, nämlich inwieweit eine Bestimmung und Interpretation des Intelligenzquotienten vertretbar oder sinnvoll ist, wird in dieser Arbeit nicht diskutiert. Der Verfasser folgt der Argumentation von Kubinger und Wurst (2000), nachdem dies nicht so ist (vgl. auch Holocher-Ertl, Kubinger & Hohensinn, 2006, 2008); dem kann der Verfasser nichts hinzufügen. Jedoch sind die Argumente und Ergebnisse, die in *dieser* Arbeit vorgestellt werden, unabhängig von dieser ersten Frage bedeutsam, weshalb auf eine weitere Darstellung derselben verzichtet werden konnte.

Natürlich trifft die Kritik, die in dieser Einleitung ansatzweise formuliert wird, nicht nur den HAWIK-IV (und seine Vorgänger), sondern auch andere (Intelligenz-)Tests. Aber sowohl die beanspruchte Gültigkeit des HAWIK-IV, als auch die große Änderungsresistenz gegenüber der Kritik an den früheren Versionen (vgl. z.B. Kubinger, 1983, 2006; Steuer, 1988) lässt ihn als ein Verfahren erscheinen, das kritisch zu beleuchten notwendig ist.

Die Bedeutung des Themas „Hochbegabungsdiagnostik“ für die vorliegende Arbeit muss allerdings relativiert werden: Weder wurde eine ausführliche Darstellung dieses Themas und der einschlägigen Theorien angestrebt, noch soll sich die kritische Auseinandersetzung mit dem HAWIK-IV einzig auf Fragen der Hochbegabungsdiagnostik beziehen. Vielmehr werden an diesen exemplarisch die Vorteile und Probleme des Einsatzes des HAWIK-IV in der Einzelfalldiagnostik diskutiert und untersucht. Dem liegt auch der Umstand zugrunde, dass Studien zur Hochbegabung den praktischen Rahmen darstellten, in dem die empirischen Untersuchungen zu dieser Arbeit durchgeführt wurden.

## **II THEORETISCHER TEIL**

### **1 DIAGNOSTIK INTELEKTUELLER HOCHBEGABUNG**

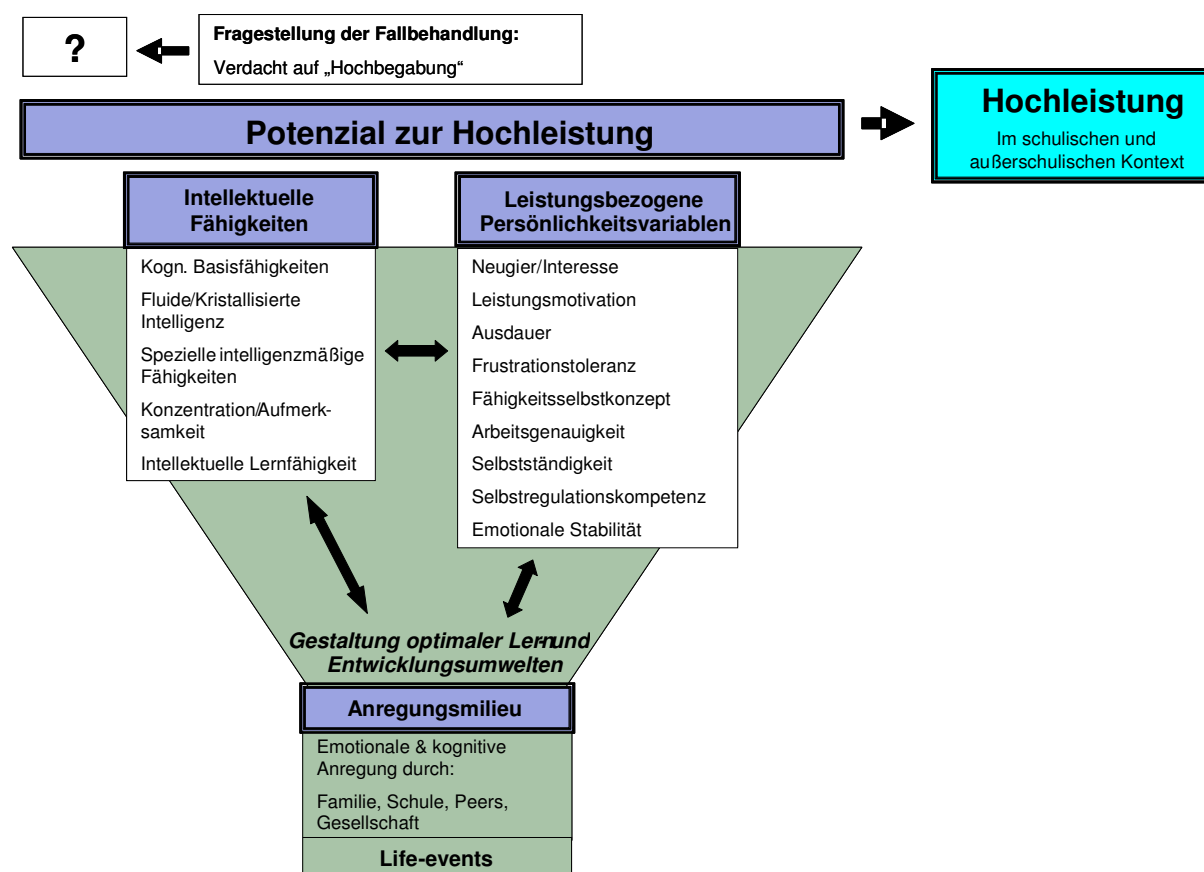
Zum Thema „intellektuelle Hochbegabung“ existieren eine Vielzahl methodischer Zugänge und Erklärungsmodelle, die im Rahmen dieser Arbeit nicht erschöpfend behandelt werden können. Vielmehr beschränkt sich diese Arbeit schon durch ihre Thematik nur auf Erklärungsmodelle bzw. Definitionen, in denen die Ergebnisse einer Intelligenz-Test-Batterie, wie es der HAWIK-IV darstellt, von Bedeutung sind. Im Rahmen dieser Modelle wird intellektuelle Hochbegabung als Disposition zu sehr hohen (kognitiven) Leistungen verstanden. Für das Diagnostizieren dieser Disposition eignen sich vor allem mehrdimensionale Modelle, in denen kognitive Fähigkeiten einen wesentlichen Faktor neben anderen wichtigen Dimensionen darstellen. Zu diesen gehört das *Wiener Diagnosemodell zum Hochleistungspotenzial* (Holocher-Ertl, Kubinger & Hohensinn, 2006, 2008), auf das sich diese Arbeit schon allein aus pragmatischen Gründen bezieht. Denn die empirischen Ergebnisse dieser Arbeit wurden in der Einzelfallbehandlung an der Test- und Beratungsstelle des Arbeitsbereichs Diagnostik der Fakultät für Psychologie an der Universität Wien gewonnen. Eine kritische Würdigung dieses Modells und eine Argumentation, weshalb dieses Modell vom Verfasser vorgezogen wird, würden den Umfang dieser Arbeit sprengen.

Eine Erörterung, inwiefern der HAWIK-IV im Rahmen ganz anderer Modelle einsetzbar wäre, wird in der vorliegenden Arbeit, mit einer Ausnahme, nicht vorgenommen. Es sind nämlich auch Hochbegabungs-Definitionen verbreitet, die sich fast ausschließlich auf die Höhe der Intelligenz beziehen, wobei damit zumeist ein Intelligenzmaß gemeint ist, das die kognitiven Fähigkeiten *global* beschreibt (meistens der Intelligenzquotient IQ). Diese Definition stützt sich zumeist auf einen Cut-off-Wert von  $IQ \geq 130$ , was den obersten 2,2 % der Intelligenzverteilung in der Referenzpopulation entspricht (vgl. Holling & Kanning, 1999). Dies geschieht trotz der Argumente, die gegen eine solche Definition sprechen, wie sie von Holocher et al. (2006, 2008) ausgeführt wurden, einzig mit der Begründung, dass es aufgrund der Verankerung dieser Definition im Schulwesen und in der Bevölkerung in Einzelfällen notwendig erscheinen mag, Klienten anhand dieses Cut-Off-Wertes als „im traditionellen Sinne hochbegabt“ zu klassifizieren.

## 1.1 Anforderungen des „Wiener Diagnosemodells zum Hochleistungspotenzial“ an ein Verfahren zur Intelligenzmessung

Das *Wiener Diagnosemodell zum Hochleistungspotenzial* ist ein psychologisch-diagnostisches Modell, das sich an den Anforderungen der Einzelfall-Begabungsdiagnostik orientiert, wie sie als Grundlage der Beratung dient. Das Diagnosemodell steht in der Tradition anderer mehrdimensionaler Modelle, die zwischen Hochbegabung als Disposition und Hochleistung als beobachtbares Verhalten unterscheiden, was dadurch explizit ausgedrückt wird, dass es der Diagnose von *Hochleistungspotenzial* dient.

**Abbildung 1.1:** Das Wiener Diagnosemodell zum Hochleistungspotenzial (Holoher-Ertl et al., 2008, S. 101)



Wie aus Abbildung 1.1 ersichtlich, basiert die Diagnose „Potenzial zur Hochleistung“ auf der Feststellung der *intellektuellen Fähigkeiten*, der *leistungsbezogenen Persönlichkeitsvariablen* und des *Anregungsmilieus*, wobei durchaus im Sinne eines Kompensationsmodells davon ausgegangen wird, dass individuelle Schwächen durch Stärken in anderen Bereichen wenigstens teilweise kompensiert werden können. Im Gegensatz dazu wird aber der impliziten Annahme, die der Bestimmung eines Gesamt-



Intelligenz-Quotienten zugrundeliegt, nämlich dass eine Kompensation kognitiver Stärken und Schwächen als „additiv“ anzunehmen wäre, aber explizit widersprochen (Holocher et al, 2006, vgl. auch Kubinger & Wurst, 2000, Kubinger 2006). Daher ist im Rahmen der Einzelfalldiagnostik die Anwendung einer solchen Testbatterie unerlässlich, die eine Profilinterpretation ermöglicht.

Die im Modell dargestellten intellektuellen Fähigkeiten beinhalten *kognitive Basisfähigkeiten*, *fluide und kristalline Intelligenz*, *Konzentration/Aufmerksamkeit* und *intellektuelle Lernfähigkeit*. Unter kognitiven Basisfähigkeiten versteht man im allgemeinen Teilleistungen, die komplexen Leistungen (wie Wahrnehmen, Lernen) zugrunde liegen, wofür aber keine verbindliche Systematik existiert. Darunter sind Leistungen, wie Differenzierungsfähigkeit, Serialität, Raum-Lage-Orientierung, unmittelbare Merkfähigkeit zu verstehen (Kubinger & Wurst, 2000), wobei die Abgrenzung zu speziellen *intelligenzmäßigen Fähigkeiten* nicht klar zu treffen ist. Unter letzteren versteht man im Wesentlichen meist komplexere Leistungen – beispielsweise rechnerische Fähigkeiten oder Raumvorstellung. Die Benennung von *fluider und kristalliner Intelligenz* verweist auf die Notwendigkeit der Diagnostik sowohl der Fähigkeiten, neue Aufgaben weitgehend unabhängig von früheren Lernerfahrungen zu lösen, als auch der Fähigkeiten, die sich als Ergebnis aller früheren Lernerfahrungen verfestigt (kristallisiert) haben. Sowohl für *Konzentration* als auch für *Aufmerksamkeit* liegen keine allgemein akzeptierten Definitionen vor. Im Wesentlichen bezeichnen sie die Fähigkeit, eine oder mehrere ausgewählte Handlungen mit ausreichender Genauigkeit auszuführen, ohne sich dabei durch irrelevante Dinge ablenken zu lassen, wobei weiters zwischen geteilter, selektiver und Daueraufmerksamkeit unterschieden wird (Amelang et al., 2006; Kubinger, 2006). Beide Begriffe werden auch im Wiener Diagnosemodell nicht genauer operationalisiert; da von diesen Faktoren aber angenommen wird, dass sie einen wesentlichen Einfluss auf die Fähigkeit zum Erbringen von Hochleistungen haben, scheinen Holocher et al. (2006) darunter die Gesamtheit der mit diesen Begriffen bezeichneten Fähigkeiten zu verstehen. Auch *intellektuelle Lernfähigkeit* wird im Rahmen des Wiener Modells nicht genauer spezifiziert; diese Fähigkeit wäre im Wesentlichen durch die Messung oder Beobachtung von Leistungssteigerungen infolge von Lernprozessen während der Vorgabe von Leistungstests zu prüfen (Kubinger, 2006).

Die Anforderungen dieses Diagnosemodells an die verwendete Intelligenztestbatterie beschränken sich aber nicht nur auf eine möglichst umfassende, reliable und valide Feststellung der oben beschriebenen kognitiven Fähigkeiten: auch *leistungsbezogene Persönlichkeitsvariablen* können und sollen durch Verhaltensbeobachtungen während der Testung erfasst werden; explizit wird diese Möglichkeit anhand des Beiblatts „Arbeitshaltungen“ des AID 2 (*Adaptives Intelligenz Diagnostikum – Version 2.1*, Kubinger & Wurst, 2000) erörtert. Als leistungsbezogene Persönlichkeitsvariablen werden aufgezählt: Neugier/Interesse, Leistungsmotivation, Ausdauer, Frustrationstoleranz, Fähigkeitsselbstkonzept, Arbeitsgenauigkeit, Selbstständigkeit, Selbstregulationskompetenz und emotionale Stabilität.

Da sich sowohl die intellektuellen Fähigkeiten als auch die Persönlichkeitseigenschaften eines Kindes nur in einer entsprechenden Entwicklungsumwelt entfalten können, gilt es im Rahmen der Einzelfalldiagnostik auch das familiäre, schulische und weitere Umfeld hinsichtlich der emotionalen und kognitiven Anregung zu untersuchen (Holocher et al., 2008). Schon durch die Betrachtung der Vielzahl möglicher Einflussfaktoren auf das Leistungsverhalten und auf die zukünftige Entwicklung weiterer Fähigkeiten und leistungsförderlicher Persönlichkeitsanteile wird deutlich, dass eine Hochbegabungsdiagnostik, die ausschließlich auf Intelligenzdiagnostik beruht, zu kurz greift. Abgesehen von der Beachtung der nicht-intellektuellen Faktoren erfordert das *Wiener Modell* eine Intelligenzdiagnostik, welche die hinreichend genaue Abschätzung einzelner Stärken und Schwächen im Sinne einer Profilinterpretation ermöglicht und nicht nur ein Gesamtergebnis (IQ); einerseits ist dies aufgrund der förderdiagnostischen Orientierung der Beratungspraxis notwendig, andererseits deshalb, weil eben *nicht* angenommen wird, dass es innerhalb realer Leistungssituationen zu einer einfachen Kompensation einer Schwäche (beispielsweise in der visuellen Differenzierungsfähigkeit) durch eine Stärke in einem anderen Bereich (beispielsweise in der akustischen Merkfähigkeit) im Sinne einer „Durchschnittsgröße“ kommt, obwohl dies viele Intelligenztest-Batterien zur Berechnung eines Gesamt-IQs implizit annehmen.

### **III EMPIRISCHER TEIL**

## **2 ZIELSETZUNG UND BESONDERHEITEN DER EMPIRISCHEN UNTERSUCHUNG**

Die vorliegende Arbeit beschäftigt sich kritisch mit der Testkonstruktion und untersucht die Eignung der Intelligenztestbatterie HAWIK-IV (*Hamburger Wechsler Intelligenztest für Kinder – IV. Übersetzung und Adaptation der WISC-IV von David Wechsler. Petermann & Petermann, 2007*) für die Hochbegabungsdiagnostik nach dem *Wiener Diagnosemodell zum Hochleistungspotenzial* und nach der Klassifikation anhand eines Cut-Off-Wertes, exemplarisch am Intelligenzquotienten. Diese Einschätzung der Eignung wird einerseits aufgrund (test)theoretischer Überlegungen, andererseits auf Basis der empirischen Daten und Erfahrungen getroffen. Da theoretische Überlegungen und empirische Datenauswertung in mehreren Bereichen ineinandergreifen, sodass Bestandteile dieser Arbeit weder eindeutig dem Theorieteil noch dem empirischen Teil zuzuordnen sind, schien es dem Verfasser nicht sinnvoll, an der strengen Trennung zwischen diesen Teilen festzuhalten. Die theoretischen Einführungen in die empirisch untersuchten Fragestellungen werden daher auch im folgenden empirischen Teil dargestellt.

Weiters muss klargestellt werden, dass es sich bei den beschriebenen Bereichen um eine nachträgliche Auswahl handelt: So wurden vom Verfasser Erfahrungen und Daten mit dem HAWIK-IV gesammelt, wobei sämtliche Bereiche der Gütekriterien eines psychologisch-diagnostischen Verfahrens im Blickfeld lagen. Erst nach der empirischen Untersuchung wurden, um den Rahmen einer Diplomarbeit nicht zu sprengen, einige Bereiche ausgewählt, welche dem Verfasser zur *Darstellung* des HAWIK-IV als *notwendig* und hinsichtlich der *Eignung* des HAWIK-IV für die Hochbegabungsdiagnostik als besonders *kritisch* erschienen.

### **2.1 Beschreibung der empirischen Untersuchung**

Die zur vorliegenden Arbeit führende empirische Untersuchung erfolgte im Rahmen einer umfassenderen Studie, die weitere Fragestellungen innerhalb des Themas „Hochbegabung bei Kindern und Jugendlichen“ untersuchte, deren Stichprobe sich aus insgesamt 49 Kindern und Jugendlichen zusammensetzte. Alle Kinder und Jugendlichen waren schon früher (zwischen 2004 und 2007) Klienten der *Test- und Beratungsstelle*

des Arbeitsbereichs Psychologische Diagnostik der Fakultät für Psychologie an der Universität Wien gewesen, wobei es in den allermeisten Fällen dezidiert um die Fragestellung „Hochbegabung“ ging. Diese Personen wurden im Zeitraum von Herbst 2007 bis Sommer 2008 im Rahmen einer Katamnesestudie neuerlich an zwei Testterminen untersucht, wobei alle Testpersonen den AID 2 (*Adaptives Intelligenz Diagnostikum – Version 2.1*, Kubinger & Wurst, 2000) bearbeiteten. Von diesen wurden 41 Personen auch mit dem HAWIK-IV getestet, der ja Thema der vorliegenden Arbeit ist; diese 41 Personen stellen die Stichprobe der vorliegenden Arbeit dar. Außerdem wurden mit allen Personen der Stichprobe ausführliche Anamnesegespräche geführt und über Elternfragebögen und den Einsatz von Persönlichkeitsfragebögen relevante Informationen eingeholt.

#### 2.1.1 Zusammensetzung der Stichprobe

Die Stichprobe setzt sich aus 41 Kindern und Jugendlichen (34 männlich und 7 weiblich) zusammen. Dieses unausgewogene Geschlechterverhältnis entspricht dem üblichen Geschlechterverhältnis der Klienten der Test- und Beratungsstelle.

Das Durchschnittsalter betrug 11;6 Jahre, die jüngste Testperson war 7;10 Jahre, die älteste 16;4 Jahre alt.

#### 2.1.2 Die Schulsituation der Testpersonen

Drei der weiblichen Testpersonen gingen in die Volksschule, ihr Notendurchschnitt im letzten Jahr betrug 1,0.

Vier weitere Mädchen besuchten die AHS, bei einem mittleren Notendurchschnitt von 1,8.

Von den männlichen Testpersonen besuchten 10 Buben die Volksschule; ihr mittlerer Notendurchschnitt lag bei 1,2.

22 Jungen besuchten eine AHS und 2 weitere eine BHS (2) mit einem Notendurchschnitt von 2,1.

## 2.2 Die verwendeten psychologisch-diagnostischen Verfahren und weitere erhobene Variablen

Nicht alle der im Rahmen der Gesamtstudie erhobenen Variablen finden in die vorliegende Arbeit Eingang, da sie Untersuchungen dienten, die im Rahmen anderer Arbeiten diskutiert wurden (beispielsweise in: Schubhart, 2008).

Es wurden folgende Verfahren angewendet:

- **AID 2** (*Adaptives Intelligenz Diagnostikum* – Version 2.1, Kubinger & Wurst, 2000)
- **HAWIK-IV** (*Hamburg-Wechsler-Intelligenztest für Kinder – IV, Übersetzung und Adaption der WISC-IV von David Wechsler*. Petermann & Petermann, 2007)

je nach Alter entweder

- **CFT 1** (*Grundintelligenztest Skala 1*, Catell, Weiß & Osterland, 1997)  
(5 Testpersonen)

oder

- **CFT 20-R** (*Grundintelligenztest Skala 2 – Revision*, Weiß, 2006), 1.Teil  
(36 Testpersonen)

und in 30 Fällen der

- **PFK 9-14** (*Persönlichkeitsfragebogen für Kinder zwischen 9 und 14 Jahren*, Seitz & Rausche, 2004) (30 Tp)

Darüber hinaus wurden weitere **leistungsbezogene Persönlichkeitsvariablen** erhoben (vgl. Schubhart, 2008). So wurden bei allen Personen sowohl die aus Anamnese und Elternfragebogen vorhandenen Informationen zum Verhalten in der Schule und zu Hause, als auch die Verhaltensbeobachtungen während der Testungen per Ratingverfahren der Testleiter zu Werten in verschiedenen leistungsbezogenen Persönlichkeitsvariablen zusammengefasst, wobei diese Variablen jeweils dreikategoriell verrechnet wurden (unterdurchschnittlich, durchschnittlich, überdurchschnittlich). In diese Ratings flossen bei den 30 Testpersonen, die ihn bearbeiteten, auch die Ergebnisse des PFK 9-14 ein.

Zu diesen leistungsbezogenen Persönlichkeitsvariablen zählen u.a.: Leistungsmotivation in der Testsituation; Ausdauer in der Testsituation; Konzentration im schulischen und familiären Kontext; Selbstüberzeugung.

### 3 VALIDITÄT

Die Validität eines Tests beschreibt das Maß, in dem er tatsächlich jenes Persönlichkeitsmerkmal misst, welches er zu messen behauptet. Es werden üblicherweise drei unterschiedliche Konzepte der Validität unterschieden (vgl. Kubinger, 2006):

1. die *inhaltliche Gültigkeit*, die gegeben ist, wenn der Test quasi selbst das zu messende Merkmal repräsentiert, wie es Schularbeiten oder andere Leistungstests in der Schule darstellen.
2. Unter *Kriteriumsvalidität* ist das Maß zu verstehen, in dem die Ergebnisse eines Tests mit einem als relevant angesehenen Außenkriterium übereinstimmen.
3. Ein anderes Konzept stellt das der *Konstruktvalidität* dar, womit gemeint ist, dass ein Test die theoriegeleiteten Annahmen über Merkmale eines psychologischen Konstruktes erfüllt. Dies zielt einerseits auf die psychologische und inhaltliche Analyse der Eigenschaften und Fähigkeiten ab, die einem Test zugrundeliegen. Andererseits erfordert die Prüfung der Konstruktvalidität das Erbringen empirischer Nachweise. Dafür gibt es wieder verschiedene statistische Ansätze, wobei beim HAWIK-IV der klassische Ansatz über die Faktorenanalyse vorgenommen wurde: Dieser Ansatz kann am besten an einem Beispiel dargestellt werden: Eine größere Anzahl von Untertests, die aufgrund inhaltlicher psychologischer Überlegungen gestaltet wurden, um das Konstrukt Intelligenz in verschiedenen Facetten abzubilden, wird einer Faktorenanalyse zugeführt. Ziel dabei ist es, möglichst wenige Faktoren (das sind beispielsweise Fähigkeitsbereiche wie „Sprachverständnis“) zu identifizieren, die dennoch möglichst viele der unterschiedlichen Ergebnisse in den Untertests erklären bzw. abbilden. Dazu wird üblicherweise eine Vorgangsweise gewählt, die dazu führt, dass jeweils mehrere Untertests im besten Fall nur einem Faktor „zugeordnet“ werden können. Eine genauere Darstellung der Methode und der häufigen Fehlinterpretationen wird hier nicht gegeben. Festzuhalten ist aber, dass diese Vorgangsweise (abgesehen von der Stichprobenabhängigkeit der Daten) in einem hohen Maß von Entscheidungen des Testkonstruktors abhängt; deshalb wäre es ein Trugschluss anzunehmen, dass alleine der Umstand, dass ein Untertest hoch auf einen bestimmten Faktor „lädt“ (also hoch mit diesem korreliert), die alleinige „Zuordnung“ dieses Untertests zu diesem Faktor rechtfertigt, wie dies

aufgrund ungenauer Darstellungen oder schlichtweg falscher Verrechnungsmodi öfters – und auch beim HAWIK-IV – vorkommt. (Kubinger, 2006, Lienert & Raatz, 1998).

### **3.1 Intelligenztheoretische Fundierung der HAWIK-IV**

Der HAWIK-IV basiert wie alle bisherigen Intelligenzskalen von Wechsler auf einer pragmatischen Beschreibung von Intelligenz als „Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen“ (Wechsler 1944, zitiert nach: Petermann & Petermann, 2007, S. 22). Die Zusammensetzung der Untertests basiert demnach weniger auf einer klar umschriebenen Intelligenztheorie als auf einer pragmatischen Zusammenstellung aufgrund klinischer und pädagogischer Erfahrungen. Daseking, Petermann und Petermann (2007) weisen aber darauf hin, dass in der Entstehung des WISC-IV (*The Wechsler Intelligence Scale for Children-Fourth Edition*, 2003), deren deutschsprachige Adaptation der HAWIK-IV ist, deutliche Bezüge zum Intelligenzmodell von Carroll, Horn und Catell (CHC-Modell) zu erkennen sind, auch wenn dies im Manual zum WISC-IV nicht explizit angegeben ist.

Das CHC-Modell postuliert drei hierarchische Ebenen der Intelligenz, wobei die zweite Ebene aus 10 Faktoren besteht, denen wiederum mehr als 70 Fähigkeiten zugrundeliegen.<sup>4</sup> Die zehn Faktoren sind: Fluide Intelligenz, Mengen- und Zahlenwissen, Kristalline Intelligenz, Lesen und Schreiben, Kurzzeitgedächtnis, Visuelle Wahrnehmung, Langzeitgedächtnis, Verarbeitungsgeschwindigkeit und Reaktionszeit/Entscheidungszeit (Daseking et al, 2007).

Nach Daseking et al. (2007, S.252) „kann davon ausgegangen werden, dass den Autoren des Testes [WISC-IV] die entsprechenden Forschungsergebnisse [des CHC-Modells] vorgelegen haben. [...] Insbesondere der Verzicht auf die Zuordnung der Untertests zu Verbal- und Handlungsteil und die aktuelle Fokussierung auf die vier Index-Werte (Faktoren) lassen darauf schließen, dass das CHC-Modell einen wesentlichen Einfluss auf die Testüberarbeitung gehabt hat.“ Diese Vermutungen werden dadurch gestützt, dass gezeigt werden konnte, dass die vier Indizes des WISC-IV fünf der oben genannten zehn CHC-Faktoren der zweiten Ebene entsprechen, nämlich den Faktoren Kristalline

---

<sup>4</sup> in anderen Versionen wurden teilweise andere oder mehr solcher Faktoren und Einzelfähigkeiten formuliert



Intelligenz, Kurzzeitgedächtnis, Visuelle Wahrnehmung, Fluide Intelligenz und Verarbeitungsgeschwindigkeit (Daseking et al., 2007). Trotz dieser Darstellung einer anzunehmenden impliziten intelligenztheoretischen Fundierung ist festzuhalten, dass der WISC-IV im Wesentlichen auf einer *faktorenanalytisch* festgelegten internen Struktur der vorhandenen Untertests basiert; diese Struktur wurde in der deutschsprachigen Stichprobe für den HAWIK-IV wiederum faktorenanalytisch belegt (vgl. Petermann & Petermann, 2007).

## **3.2 Konstruktvalidität**

### **3.2.1 Die Faktorenstruktur des HAWIK-IV**

Die Faktorenstruktur des WISC-IV (und damit des HAWIK-IV) ergibt sich aus einer 4-Faktoren-Lösung einer Faktorenanalyse aller 15 Untertests. Die vier Faktoren und ihre entsprechenden Indizes sind: *Sprachverständnis (SV)*, *Wahrnehmungsgebundenes Logisches Denken (WLD)*, *Arbeitsgedächtnis (AGD)* und *Verarbeitungsgeschwindigkeit (VG)*. Diese Indizes werden in der Testpraxis durch 10 Kerntests gebildet (Petermann & Petermann, 2007).

Der Vier-Faktoren-Lösung wurde in der deutschsprachigen Variante der Vorzug gegenüber einer 5-Faktoren-Lösung gegeben, da diese nur unwesentlich mehr Varianz erkläre; auf diesen fünften Faktor lädt der Untertest *Rechnerisches Denken*. In der letztlich beibehaltenen 4-Faktoren-Lösung wird dieser (optionale) Untertest dem Index *Arbeitsgedächtnis* zugerechnet (Daseking, Petermann & Petermann, 2007).

Innerhalb der 4 resultierenden Indizes korrelieren die Untertests des Index *Sprachverständnis* am höchsten miteinander: Es sind dies die Kerntests *Gemeinsamkeiten Finden (GF)*, *Wortschatz-Test (WT)* und *Allgemeines Verständnis (AV)*.

Die Untertests des Index *Wahrnehmungsgebundenes Logisches Denken* korrelieren laut Manual sowohl untereinander als auch mit den sprachlichen Untertests hoch und auch mit den Untertests des Index *Arbeitsgedächtnis* mittelmäßig. Es sind dies die Kerntests *Mosaik-Test (MT)*, *Bildkonzepte (BK)* und *Matrizen-Test (MZ)*.

Die Kerntests zum *Arbeitsgedächtnis*, das sind *Zahlen nachsprechen (ZN)* und *Buchstaben-Zahlen-Folgen (BZF)* korrelieren am höchsten miteinander und mit den Untertests des Index *Sprachverständnis*.

Die Kerntests des Index *Verarbeitungsgeschwindigkeit* sind der *Zahlen-Symbol-Test (ZST)* und die *Symbol-Suche (SYS)*. Diese korrelieren in erster Linie untereinander (Petermann & Petermann, 2007).

Der HAWIK-IV beinhaltet auch optionale Untertests, die entweder anstelle eines der Kerntests zur Berechnung der Indexwerte herangezogen werden können oder zusätzlich vorgegeben werden können, ohne in die Berechnung der Indizes und des Gesamt-IQs einzugehen. Das sind die Untertests *Rechnerisches Denken (RD)* für den Index *Arbeitsgedächtnis*, *Allgemeines Wissen (AW)* und *Begriffe Erkennen (BEN)* für den Index *Sprachverständnis*, *Bilder Ergänzen (BE)* für den Index *Wahrnehmungsgebundenes Logisches Denken* und der *Durchstreich-Test (DT)* für den Index *Verarbeitungsgeschwindigkeit* (Petermann & Petermann, 2007). Mangels entsprechender Daten wird auf die letzten drei genannten Untertests in dieser Arbeit nicht eingegangen.

### 3.2.2 Kritik an der Index- Zuordnung des Untertest Rechnerisches Denken

Die Zuordnung des optionalen Untertest *Rechnerisches Denken (RD)* zum Index *Arbeitsgedächtnis* erscheint problematisch. Wie aus dem Manual hervorgeht, lädt dieser Test bei einer 5-Faktoren-Lösung auf einen eigenen Faktor, der allerdings so wenig zusätzliche Varianz erklärt, dass die 4-Faktorenlösung vorgezogen wurde.

Dass dieser Untertest demnach definitionsgemäß nicht „rechnerische Fähigkeiten“, oder, wie es im CHC-Modell heißt, „Mengen- und Zahlenwissen“ prüft, sondern dem Index *Arbeitsgedächtnis* zugerechnet wird, erscheint inhaltlich fragwürdig. Dass er darüber hinaus sogar standardmäßig nicht vorgegeben wird und auch nicht in den Gesamt-IQ einfließt, ist aus förderdiagnostischer Sicht bedenklich (Holocher et al., 2008). Prägnant könnte man nämlich fragen: wie soll Schullaufbahnberatung stattfinden, ohne Wissen um die rechnerischen Fähigkeiten eines Kindes?

Darüber hinaus ist die Zuordnung zum Index *Arbeitsgedächtnis* auch in Zusammenhang mit der postulierten intelligenztheoretischen Fundierung des HAWIK-IV durch das CHC-Modell einigermaßen unverständlich: Denn unter der Annahme des CHC-Modells,

dass *Mengen- und Zahlenwissen* ein eigenständiger, die Gesamtintelligenz fundierender Faktor ist, ist der faktorenanalytische Befund, dass *Rechnerisches Denken* auf einen eigenen Faktor lädt, doch ein starkes Argument, diesen Untertest auch genau so zu interpretieren, auch wenn das einem der Ziele einer Faktorenanalyse – nämlich der Reduktion der Dimensionen – vordergründig widerspricht.

Dies geschieht im HAWIK-IV aber nicht, und zwar mit der Begründung, dieser Faktor würde nur unzureichend zusätzliche Varianz erklären. Dies ist aber kein zutreffendes Argument, wie sich anhand von einer Überlegung zum Vorgehen der Faktorenanalyse zeigen lässt: Angenommen, es gibt wirklich nur einen einzigen Untertest, der *Mengen- und Zahlenwissen* erfasst, so ist es in einer Faktorenanalyse ja gar nicht möglich, dass sich ein entsprechender Faktor *Mengen- und Zahlenwissen* extrahieren lässt, der mehr Varianz erklärt, als eben dieser eine Untertest einbringt.

Zwar ist durchaus anzunehmen, dass die Fähigkeit *Mengen- und Zahlenwissen* in geringem Maße auch in andere Untertests einfließt. Dennoch gilt: wenn zu jedem anderen Index zumindest zwei oder drei Untertests existieren, die ähnliches messen, und deswegen auch (hoch) positiv miteinander korrelieren, so werden die daraus resultierenden Faktoren fast zwangsläufig mehr Varianz erklären als ein Faktor, auf den (im Wesentlichen) nur ein einziger Untertest lädt. Da die Aufgaben des Untertests *Rechnerisches Denken* der Testperson nur einmal vorgelesen, also gemerkt werden müssen, ist durchaus zu erwarten, dass ein Teil der Varianz im *Rechnerischen Denken* durch einen Faktor, der Merkfähigkeit abbildet, erklärt wird. Dies vermindert aber die Varianz, die ein zusätzlicher Faktor *Mengen- und Zahlenwissen* erklären könnte. Dies gilt selbst dann, wenn die Leistungen im *Rechnerischen Denken* abgesehen davon das Konstrukt *Mengen- und Zahlenwissen* perfekt abbilden würden.

Diese Überlegungen führen zu dem Schluss, dass das Argument, ein fünfter Faktor würde nicht genug Varianz erklären, einzig im Sinne der angestrebten Dimensionsreduktion seine Berechtigung hat: ist der Sinn der Faktorenanalyse nämlich der, aus einer Menge an Variablen *möglichst wenig* Faktoren zu extrahieren, die möglichst viel Varianz erklären, dann ist es natürlich sinnvoll, keine Faktoren mit einzubeziehen, die nur wenig Varianz erklären. Diese Vorgangsweise dann aber so zu interpretieren, dass alle Untertests in inhaltlich sinnvoller Weise anhand der resultierenden Faktoren erklärt werden könnten, geht über die Aussagekraft einer

Faktorenanalyse hinaus, insbesondere dann, wenn inhaltliche Überlegungen dagegen sprechen.

Betrachtet man dieses Thema - losgelöst von faktorenanalytischen Ergebnissen – nämlich inhaltlich, so lässt sich konstatieren: Anhand der Darstellung einiger Items lässt sich klar nachvollziehbar machen, dass der Untertest *Rechnerisches Denken* neben der akustischen Merkfähigkeit zumindest noch eine andere Fähigkeit erfasst, nämlich die zu *rechnen*.<sup>5</sup>

Zur Darstellung sind hier als Beispiele die drei Items mit relativen Lösungshäufigkeiten nahe 0,5 und außerdem das Item (Nr. 34) mit der geringsten Lösungshäufigkeit in der Stichprobe angeführt (zitiert nach: Petermann & Petermann, 2007: S. 334 f.).

Item 26: „Eine Schule hat in jedem Klassenzimmer 25 Schüler. Wie viele Klassenzimmer hat die Schule, wenn sie insgesamt 500 Schüler hat?“

Item 28: „Eine Familie fuhr drei Stunden lang mit dem Auto, machte dann eine Pause und fuhr dann noch zwei Stunden lang weiter. Sie fuhren insgesamt 300 Kilometer weit. Mit welcher Durchschnittsgeschwindigkeit war die Familie gefahren?“

Item 29: „Anja kauft sich ein gebrauchtes Fahrrad [...] für zwei Drittel vom Neupreis. Sie musste 20 Euro [...] dafür bezahlen. Wie viel kostete das Fahrrad neu?“

Item 34: „Philipp geht eine Stunde früher nach Hause als Sven. Philipp fährt 40 km/h, und Sven fährt 60 km/h. Beide fahren in dieselbe Richtung. Wie viele Kilometer Vorsprung hat Sven fünf Stunden, nachdem Philipp losgefahren ist?“

Dass diese Items Rechenfähigkeit und nicht nur Merkfähigkeit prüfen, bedarf bei Betrachtung der Beispiele keiner weiteren Argumentation. Rechenaufgaben jedoch vorzugeben, *nur* um Merkfähigkeit zu erheben, ist weder valide (da schlechte Rechner

---

<sup>5</sup> Der Untertest *Rechnerisches Denken* weist in der aus nur 24 Personen bestehenden (und relativ homogenen) Stichprobe, die diesen optionalen Untertest bearbeiteten eine Korrelation von  $r = 0,56$  bzw.  $\rho = 0,59$  (Spearman) zum AID 2 – Untertest *Angewandtes Rechnen* auf, was im Vergleich zu anderen Korrelationen zwischen „analogen“ Untertests des HAWIK-IV und des AID 2 sehr hoch ist. Da im AID 2 die Aufgaben schriftlich vorgelegt werden, also unabhängig von einer verbal-akustischen Merkleistung zu lösen sind, spricht dies deutlich dafür, dass die gemeinsame Varianz in beiden Untertests die Fähigkeit zu *rechnen* wiedergibt.

selbst bei guter Merkfähigkeit nicht auf die Lösung kommen) noch ökonomisch, und für schlechte Rechner auch kaum zumutbar!

Aus intelligenztheoretischen und förderdiagnostischen Überlegungen scheint das Prüfen des Mengen- und Zahlenwissens aber durchaus sinnvoll und notwendig. Wenn die Rechenaufgaben jedoch so stark mit Merkfähigkeit konfundiert sind, dass sie nicht eindeutig interpretiert werden können, dann bedarf es wohl anderer Items oder anderer Vorgaberegeln (etwa durch Vorlegen der Angabe, wie es im AID 2 realisiert wurde), damit es möglich wird, die Rechenfähigkeit eines Kindes möglichst unabhängig von ihrer Merkfähigkeit zu erfassen. Andernfalls liefern sie kaum zusätzliche diagnostisch verwertbare Informationen.

### **3.3 Eigene Untersuchungen zur Validität**

Zur Untersuchung der Kriteriums- und der Konstruktvalidität wird u.a. von Lienert und Raatz (1998) vorgeschlagen, den zu untersuchenden Test gemeinsam mit mehreren Außenkriterien sowie anderen Tests, die ähnliches erfassen sollen, und außerdem Tests, die andere Merkmale erfassen sollen, einer Faktorenanalyse zuzuführen; valide im Sinne dieses Ansatzes wäre also ein Test, der gemeinsam mit den Außenkriterien und den konstruktnahen Tests auf einen Faktor hoch laden (konvergente Validität), nicht aber mit den konstruktfernen Tests (diskriminante Validität) (Kubinger, 2006). In Ermangelung eines validen Außenkriteriums musste sich die Untersuchung auf die Zusammenhänge zu anderen Tests beschränken. Als konstruktnahe bzw. -ferne Tests wurden die im Rahmen dieser Untersuchung erhobenen Werte der Intelligenztests AID 2 und CFT 20-R in die Faktorenanalyse miteinbezogen.

#### **3.3.1 Faktorenanalyse der HAWIK-IV-Tests und anderer konstruktnaher und -ferner Tests**

Die normierten Werte der zehn Kerntests des HAWIK-IV, weiters der Untertest *Rechnerisches Denken*, der CFT 20-R und alle Untertests des AID 2 (ohne Zusatztests) wurden einer Faktorenanalyse zugeführt. Die beiden T-Werte des AID 2-Untertests 5 *Unmittelbares Reproduzieren* (nämlich *vorwärts* und *rückwärts*) wurden zu einem Wert (dem arithmetischen Mittel der beiden) zusammengefasst, um sie mit dem analogen HAWIK-IV-Untertest *Zahlen Nachsprechen (ZN)* vergleichbar zu machen, der auch nur einen Wert für vorwärts und rückwärts berechnet. Der Untertest *Allgemeines Wissen*

wurde nicht aufgenommen, da nach Aufnahme dieses, nur von 24 Personen bearbeiteten Test die Korrelations- Matrix der Faktorenanalyse nicht positiv definit und damit nicht lösbar war. Die „vollständige“ Faktorenanalyse aller oben angeführter Untertests führte zu einer Sieben-Faktoren-Lösung (vgl. Tabelle 3.1; Methode: Hauptkomponentenanalyse, Extraktion aller Faktoren mit Eigenwert größer als 1, Rotation nach Varimax-Kriterium) mit 74% erklärter Varianz.<sup>6</sup>

Es zeigen sich mehrere stabile Faktoren, die relativ unabhängig davon bestehen, ob einzelne Untertests zusätzlich oder eben nicht aufgenommen wurden. Auf den **ersten Faktor** laden die HAWIK-IV-Untertests<sup>7</sup> *Allgemeines Verständnis (AV)* und *Gemeinsamkeiten Finden (GF)* und die AID 2-Untertests 1 (*Alltagswissen*), 6 (*Synonyme Finden*), 9 (*Funktionen Abstrahieren*) und 11 (*Soziales Erfassen und sachliches Reflektieren*) am höchsten (mit Faktorladungen zwischen 0,66 und 0,80); dieser Faktor wird demnach als „sprachliche Intelligenz“ gedeutet. Wird der Untertest *Allgemeines Wissen (AW)* zusätzlich aufgenommen und werden dafür die Untertests 4 und 7 des AID 2 aus der Faktorenanalyse genommen, so ist die Faktorenanalyse auch mit *Allgemeines Wissen* lösbar: in dieser resultierenden Lösung lädt auch dieser Untertest erwartungsgemäß auf diesen ersten Faktor.

Der *Wortschatz-Test (WT)*, der eigentlich auch zu den „sprachlichen“ Tests gehört, lädt auf diesen Faktor nur in zweiter Linie (mit 0,42); deutlich höher (mit 0,76) lädt er auf einen anderen Faktor, der hier erst als siebenter Faktor dargestellt wird.

Der **zweite Faktor** lässt sich als „akustische Merkfähigkeit“ oder, der Diktion des HAWIK-IV entsprechend als „Arbeitsgedächtnis“ benennen, da auf ihn ausschließlich *Zahlen nachsprechen (ZN)* und *Buchstaben-Zahlen-Folgen (BZF)* und der AID 2-Untertest 5 (*Unmittelbares reproduzieren - numerisch*), hoch laden (zwischen 0,68 und 0,86; dies zeigt sich relativ unabhängig von anderen aufgenommenen oder ausgeschlossenen Variablen). Zu erwähnen ist auch, dass sich die bereits diskutierte Zuordnung des Untertests *Rechnerisches Denken* zum Index *Arbeitsgedächtnis* durch

---

<sup>6</sup> Einschränkung ist anzumerken, dass eine der Voraussetzungen der Faktorenanalyse, nämlich Normalverteilung der Variablen in einigen Fällen nicht gegeben war.

<sup>7</sup> Zur besseren Unterscheidbarkeit werden die HAWIK-IV-Untertests mit ihrem *Namen* und die AID2-Untertests mit ihrer Nummer (mit dem *Namen* in Klammern) bezeichnet.

diese Faktorenanalyse weiter in Frage stellen lässt: dieser Untertest lädt auf diesen Faktor praktisch nicht (- 0,059).

Der AID 2-Untertest 4 (*Soziale und Sachliche Folgerichtigkeit*) lädt in der hier dargestellten 7-Faktorenlösung auf diesen zweiten Faktor negativ (-0,59), was weder aus dem Verständnis des Tests noch aus den Daten erklärt werden kann. Gleiches gilt für den HAWIK-IV-Untertest *Bildkonzepte (BK)*, der ebenfalls negativ (-0,47) auf diesen Faktor lädt.

Der **dritte Faktor** bezeichnet in erster Linie „Verarbeitungsgeschwindigkeit“, da auf ihn die HAWIK-IV-Untertests *Zahlen-Symbol-Test (ZST)* und *Symbol-Suche (SYS)* sowie der AID 2-Untertest 7 (nämlich die *Kodiermenge*) hoch (zwischen 0,72 und 0,85) laden. Außerdem lädt auch der Wert für *Assoziationen* des AID 2-Untertests 7 mit etwa 0,51 auf diesen Faktor, was sich durch die voneinander abhängigen Aufgabenstellungen von *Kodieren und Assoziieren* erklären lässt, weswegen dies auch nicht gegen die diskriminante Validität spricht (auch wenn dieser zweite Wert (Assoziieren) des Untertest 7 etwas anderes als Arbeitsgeschwindigkeit, nämlich inzidentelles Lernen, aber eben in Abhängigkeit von der Leistung beim Kodieren, prüft).

Auf den **vierten Faktor** laden die AID2-Untertests 2 (*Realitätssicherheit*), 8 (*Antizipieren und Kombinieren*) und 7 (nämlich: *Assoziationen*) relativ hoch (zwischen 0,55 und 0,78) und die HAWIK-IV-Untertests *Rechnerisches Denken (RD)*, *Matrizen-Test (MZ)* und *Mosaik-Test (MT)* mit Ladungen zwischen 0,40 und 0,45. Als gemeinsame Basis dieser Untertests scheinen visuelle Fähigkeiten wie Diskriminationsfähigkeit oder visuelle Merkfähigkeit eine Rolle zu spielen. Allerdings zeigt sich dieser Faktor als empfindlich gegenüber dem Ausschließen einzelner Variablen: auf diesen Faktor laden eigentlich nur die beiden AID 2-Untertests 2 und 8 stabil, aber keiner der HAWIK-IV-Untertests, weshalb daraus auch keine Schlüsse für den HAWIK-IV gezogen werden können.

Der **fünfte Faktor** ist wiederum relativ stabil gegenüber dem Zufügen oder Wegnehmen anderer Untertests, auf ihn laden sowohl der CFT 20-R als auch der AID2-Untertest 3 (*Angewandtes Rechnen*) und der HAWIK-IV-Untertest *Rechnerisches Denken (RD)* (zwischen 0,70 und 0,85). Er scheint damit sowohl rechnerische Fähigkeiten als auch formal-logisch analytische Denkvorgänge abzubilden.

Auf den **sechsten Faktor** laden die beiden Untertests *Mosaik-Test (MT)* und der Untertest 10 des AID2 (*Analysieren und Synthetisieren – abstrakt*), welche beide sehr ähnliche Aufgabestellungen haben, nämlich abstrakte Muster nach Vorlage mit Würfeln nachzubauen (Ladungen zwischen 0,65 und 0,82). Als Unterschied zeigt sich in der Faktorenanalyse, dass der *Mosaik-Test (MT)* auch auf den Faktor „Verarbeitungsgeschwindigkeit“ mit Ladungen zwischen 0,4 und 0,65 (je nach zusätzlich aufgenommenen Untertests) lädt, was darauf zurückzuführen ist, dass der *Mosaik-Test* Gutpunkte für rasches Lösen vergibt, was der AID 2-Untertest 10 nicht tut.

Auf den **siebenten Faktor** lädt der HAWIK-IV-Untertest *Wortschatztest (WT)* hoch (0,76), mit deutlich niedrigeren Ladungen (von 0,51 bzw. 0,49) die Untertests *Bildkonzepte (BK)* und *Matrizen-Test (MZ)*. Dieser Faktor ist nur schwer interpretierbar, zumal dies weder mit der prognostizierten Faktorenstruktur des HAWIK-IV übereinstimmt, noch inhaltlich leicht zu erklären ist: zwar gehören die Untertests *BK* und *MZ* zum selben Index, nämlich *Wahrnehmungsgebundenes Logisches Denken*, und haben als gemeinsame Basis wohl visuelle Fähigkeiten und Reasoning, dies gilt aber nicht für den *Wortschatz-Test*.

**Tabelle 3.1:** Rotierte Ladungsmatrix der Untertests des HAWIK-IV, des AID 2 und des CFT 20-R (normierte Werte); Extraktionsmethode: Hauptkomponentenanalyse, Rotationsmethode: Varimax mit Kaiser-Normalisierung. Im Text erwähnte Werte wurden **fett** gedruckt.

	1	2	3	4	5	6	7
Allgemeines Verständnis	<b>,795</b>	-,190	,047	,070	,080	-,119	,308
AID 2 UT6	<b>,778</b>	,040	,098	,265	-,192	-,006	,010
AID 2 UT1	<b>,738</b>	,123	,044	,138	,084	,265	,129
AID 2 UT9	<b>,725</b>	-,026	,255	-,014	,341	-,122	-,076
Gemeinsamkeiten Finden	<b>,691</b>	,057	,030	-,219	,054	-,035	,315
AID 2 UT11	<b>,663</b>	-,332	,273	,193	,093	,163	,000
AID 2 UT5-Mittel	,012	<b>,859</b>	,099	-,049	,090	-,141	,203
Zahlen nachsprechen	-,154	<b>,803</b>	,194	,153	,006	,099	,007
Buchstaben-Zahlen-Folgen	,268	<b>,679</b>	,009	,193	,215	,211	-,164
AID 2 UT 4	,414	<b>-,593</b>	,091	-,101	,194	,112	,427



AID 2 UT7 Kodierm.	,061	,114	<b>,848</b>	,209	-,073	-,181	,121
Zahlen-Symbol-Test	,283	,116	<b>,769</b>	-,009	-,003	,384	,017
Symbol-Suche	,211	,131	<b>,716</b>	,095	,243	,349	-,085
AID 2 UT 8	,057	,197	,167	<b>,783</b>	,159	,185	,011
AID 2 UT 7 Ass.	,075	-,155	<b>,514</b>	<b>,594</b>	,026	-,302	-,003
AID 2 UT 2	,464	,319	,014	<b>,551</b>	,066	,124	,248
AID 2 UT 3	,109	,161	,105	-,065	<b>,814</b>	,005	-,011
Rechnerisches Denken	,189	-,059	-,185	<b>,452</b>	<b>,758</b>	-,011	-,023
CFT 20-R Teil1	-,100	,071	,079	,142	<b>,641</b>	,451	,253
AID 2 UT 10	,061	,005	,015	-,067	,160	<b>,844</b>	,082
Mosaik-Test	-,016	-,085	<b>,397</b>	,396	-,160	<b>,615</b>	-,099
Wortschatz-Test	<b>,418</b>	,058	-,074	-,007	,045	,053	<b>,760</b>
Bildkonzepte	,046	<b>-,466</b>	,079	,175	-,167	,317	<b>,513</b>
Matrizen-Test	,225	,095	,246	,448	,271	-,184	<b>,494</b>

### 3.3.2 Faktorenanalyse der HAWIK-IV-Untertests

Eine weitere – nur den HAWIK-IV betreffende – Faktorenanalyse (mit den zehn Kerntest und den Untertests *Rechnerisches Denken* und *Allgemeines Wissen*) zeigt eine Vier-Faktoren-Lösung (siehe Tabelle 3.2; Methode: Hauptkomponentenanalyse, Extraktion aller Faktoren mit Eigenwert größer als 1, Rotation nach Varimax-Kriterium).

Auf den **ersten Faktor** laden die vier zum Index *Sprachverständnis* gehörigen Untertests (inkl. *Wortschatz-Test*), auf einen **zweiten Faktor** die *Verarbeitungsgeschwindigkeit* messenden Untertest *Zahlen-Symbol-Test* und *Symbol-Suche*, aber auch der *Mosaik-Test*, der ja, wie schon erwähnt, eine starke Geschwindigkeitskomponente hat. Dieser Untertest gehört aber laut HAWIK-IV-Manual zum Index *Wahrnehmungsgebundenes Logisches Denken (WLD)*, der sich in dieser Faktorenlösung nicht abbildet. Der **dritte Faktor** umfasst die beiden *Arbeitsgedächtnis* prüfenden Untertests *Zahlen nachsprechen* und *Buchstaben-Zahlen-Folgen*, nicht aber den Untertest *Rechnerisches Denken*, der ja auch zu diesem Index gezählt wird. Dieser lädt nämlich gemeinsam mit dem *Matrizen-Test*, (der laut Manual dem Index *WLD*

zugerechnet wird), auf den **vierten Faktor**. Der dritte zum Index *WLD* gehörige Untertest *Bildkonzepte* lädt nur auf einen Faktor hoch, und zwar – aber mit negativem Vorzeichen! – auf den dritten Faktor (*Arbeitsgedächtnis*), was sich einer inhaltlichen Erklärung weitgehend entzieht. Ein Herausnehmen dieses Untertests aus der Faktorenanalyse ändert an der sonst beschriebenen Faktorenstruktur nichts: Gemeinsam erklären diese Faktoren 68% der Varianz (bzw. 72% nach Herausnahme von *Bildkonzepte*)

**Tabelle 3.2:** Rotierte Ladungsmatrix der Untertests des HAWIK-IV (normierte Werte); Extraktionsmethode: Hauptkomponentenanalyse, Rotationsmethode: Varimax mit Kaiser-Normalisierung. Im Text erwähnte Werte wurden **fett** gedruckt.

	1	2	3	4
Gemeinsamkeiten Finden	<b>,851</b>	-,004	,147	-,061
Allgemeines Verständnis	<b>,775</b>	,059	-,174	,357
Wortschatztest	<b>,771</b>	-,038	-,064	,087
Allgemeines Wissen	<b>,701</b>	,243	-,223	,229
Zahlen-Symbol-Test	,251	<b>,859</b>	,095	-,013
Mosaiktest	-,139	<b>,792</b>	-,049	,003
Symbol-Suche	,071	<b>,777</b>	,212	,162
Zahlen Nachsprechen	-,056	,244	<b>,779</b>	-,076
Buchstaben-Zahlen-Folgen	,089	,249	<b>,733</b>	,312
Bildkonzepte	,252	,320	<b>-,650</b>	,018
Rechnerisches Denken	,055	-,065	,075	<b>,894</b>
Matrizen-Test	,319	,209	,040	<b>,619</b>

### 3.4 Diskussion und Zusammenfassung der Ergebnisse zur Validität

Die theoretische Fundierung des HAWIK-IV (durch das CHC-Modell) wurde der faktorenanalytische Konstruktvalidierung nachträglich zugeschrieben; wiewohl dieses Modell für die Testinterpretation bedeutsam ist, ist es nicht Grundlage der Konstruktion des HAWIK-IV. Doch eben diese Konstruktvalidierung mittels faktorenanalytischer Studien ist insofern wenig überzeugend, als Faktorenanalysen *prinzipiell* keine

eindeutigen Aussagen ermöglichen<sup>8</sup> und darüber hinaus in den vorliegenden Daten nur teilweise replizierbar waren.<sup>9</sup>

Über die im Rahmen dieser Studie durchgeführten Hauptkomponentenanalysen der HAWIK-IV-Untertests, AID 2-Untertests und des CFT 20-R lässt sich zusammenfassend sagen, dass sich eine gegenüber der Aufnahme oder dem Ausscheiden einzelner Variablen relativ stabile Faktorenstruktur von sieben Faktoren zeigt. Diese Faktoren sind: ein „sprachlicher“ Faktor (Wortschatz, verbales Schlussfolgern, sprachlicher Ausdruck sozialer und sachlicher Regeln und von Alltagswissen), je ein Faktor für Verarbeitungsgeschwindigkeit und für Arbeitsgedächtnis (bzw. akustische Merkfähigkeit), ein Faktor „rechnerische/formal-logische Intelligenz“, ein Faktor für die „Würfelaufgaben“ und ein schwer interpretierbarer Faktor, auf dem die Untertests *Wortschatz-Test (WT)*, *Bildkonzepte(BK)* und *Matrizen-Test (MZ)* laden. Ein weiterer Faktor beschreibt in erster Linie Untertests des AID 2, die keine inhaltliche Entsprechung zu Untertests des HAWIK-IV haben. Das Ergebnis stützt – insofern Faktorenanalysen das überhaupt können (vgl. Kubinger, 2006) – den Befund, dass die konstruktnahen Untertests des HAWIK-IV und des AID 2 (mit Einschränkungen hinsichtlich des *Wortschatz-Tests*) etwas Ähnliches messen, und dass die vom HAWIK-IV angenommenen Indizes *Sprachverständnis*, *Arbeitsgedächtnis* und *Arbeitsgeschwindigkeit* sich im Wesentlichen wiederfinden. Der Indizierung des HAWIK-IV-Manuals allerdings widersprechend lässt sich der Untertest *Rechnerisches Denken* im Rahmen dieser Faktorenanalysen nicht dem Faktor *Arbeitsgedächtnis* zuordnen, sondern lädt auf einen gemeinsamen Faktor mit dem AID 2-Untertest 3 (*Angewandtes Rechnen*) und dem CFT 20-R. Zumindest in dieser Stichprobe scheint der Untertest *Rechnerisches Denken* eher numerische und fluide Intelligenz zu prüfen als Arbeitsgedächtnis. Auch der Index *Wahrnehmungsgebundenes logisches Denken* lässt sich in der Faktorenstruktur nicht wiederfinden: die dazugehörigen Untertests *Mosaik-Test* einerseits und *Bildkonzepte* und *Matrizen-Test* andererseits laden fast ausschließlich auf unterschiedliche Faktoren. Auf die einzelnen Untertests bezogen zeigen sich zwar deutliche Hinweise für konvergente Validität des *Mosaik-Test* mit dem Untertest 10 des

---

<sup>8</sup> Unter anderem aufgrund der Stichprobenabhängigkeit aller korrelativen Verfahren (vgl. Kubinger, 2006).

<sup>9</sup> Mitunter scheint es sogar so, dass die errechnete Faktorenstruktur der publizierten theoretischen Fundierung durch das CHC-Modell zumindest teilweise *nicht* entspricht; als Beispiel dafür kann der Untertest *Rechnerisches Denken* gelten.

AID 2, allerdings mit der starken Einschränkung, dass der *Mosaik-Test* durch seine *Speed*-Komponente auch stark von der Arbeitsgeschwindigkeit abhängt. Außerdem zeigt sich, dass die Untertests *Matrizen-Test* und *Bildkonzepte*, die beide fluide Intelligenz messen sollen, nicht auf einen gemeinsamen Faktor mit dem CFT 20-R (prüft definitionsgemäß fluide Intelligenz) laden. Für die Untertests *Bildkonzepte* und *Matrizen-Test* zeigt sich in diesen Faktorenanalysen kein Hinweis, dass die von den Testautoren behauptete Prüfung fluider Intelligenz in wenigstens ähnlicher Art und Weise gegeben ist, wie sie der CFT 20-R leistet.

Das Ergebnis der zweiten, nur die HAWIK-IV-Untertests betreffende Faktorenanalyse stützt den oben dargestellten Befund betreffend die Untertests *Rechnerisches Denken*, *Mosaik-Test*, *Bildkonzepte* und *Matrizen-Test*: Die Zuordnung des Untertests *Rechnerisches Denken* zum Index *Arbeitsgedächtnis* ist anhand dieser Daten nicht nachvollziehbar. Auch der Index *Wahrnehmungsbezogenes Logisches Denken*, der die Untertests *Mosaik-Test*, *Bildkonzepte* und *Matrizen-Test* beinhalten sollte, lässt sich aus der Faktorenanalyse dieser Daten nicht replizieren.

Hinsichtlich der Möglichkeit einer Profilinterpretation, wie dies vom Wiener Diagnosemodell zum Hochleistungspotenzial gefordert wird, lassen sich die Ergebnisse so zusammenfassen, dass die vier Indizes des HAWIK-IV für eine differenzierte Profilinterpretation nicht ausreichen: einerseits *fehlen* wesentliche Bereiche (beispielsweise numerische Intelligenz/Rechenfähigkeit), andererseits ist beispielsweise der Index *Wahrnehmungsgebundenes logisches Denken* inhaltlich kaum eindeutig zu interpretieren; darüber hinaus war er durch die vorliegende Faktorenanalyse nicht zu replizieren. Die anderen drei Indizes (*Sprachverständnis*, *Arbeitsgedächtnis* und *Verarbeitungsgeschwindigkeit*) scheinen für eine relativ grobe Profilinterpretation sinnvoll einsetzbar.

Auf Untertestebene wären vor allem die Untertests des Index *Wahrnehmungsgebundenes logisches Denken* als auch der Untertests *Rechnerisches Denken* für eine Profilinterpretation interessant: Allerdings konnte die behauptete Prüfung fluider Intelligenz durch die Untertests *Matrizen-Test* und *Bildkonzepte* (jedenfalls im Vergleich zum CFT 1 bzw. CFT 20-R) im Rahmen der Faktorenanalysen nicht plausibel gemacht werden. Die niedrigen korrelativen Zusammenhänge zwischen diesen Test können aber auch durch Deckeneffekte oder zu wenig reliable Messungen zustande gekommen sein. Der Untertest *Mosaik-Test*, der von seiner Aufgabenstruktur

geeignet wäre, Raum-Lage-Orientierung und die Fähigkeit zur Strukturierung visuellen Materials zu prüfen, zeigt aber – aufgrund der Zeitgutpunkte – in der Faktorenanalyse eine starke Nähe zum Faktor *Verarbeitungsgeschwindigkeit*, weshalb auch hier die eindeutige Interpretation erschwert ist. Inwiefern diese drei Einzelskalen und der Untertest *Rechnerisches Denken* aber überhaupt genügend messgenau wären, um eine Profilinterpretation auf Untertestebene zu ermöglichen, wird im Kapitel 5 diskutiert.

## **4 KRITIK AN DER DREIKATEGORIELLEN VERRECHNUNG DER UNTERTESTS GF, WT UND AV – HINSICHTLICH VALIDITÄT, SKALIERUNG UND PRAKTISCHER HANDHABUNG**

### **4.1.1 Darstellung des Vorgabe- und Verrechnungsmodus**

In den drei zum Index *Sprachverständnis* zählenden Untertests *Gemeinsamkeiten finden*, *Wortschatz-Test (WT)* und *Allgemeines Verständnis (AV)* werden die Antworten dreikategoriell verrechnet, also entweder mit zwei Punkten (volle Punkteanzahl), mit nur einem Punkt (halbe Punkteanzahl für weniger treffende oder nicht vollständige Antworten) oder mit null Punkten bewertet. Dabei ist es den Personen erlaubt, spontan auch mehrere Antworten zu äußern, wobei es auch möglich ist, dass sich zwei 1-Punkt-Antworten zu einer 2-Punkt-Antwort ergänzen (bzw. eine eindeutig grob *falsche* Antwort eine andere, richtige Antwort „entwertet“). Bei einigen besonders gekennzeichneten Null-Punkt- oder 1-Punkt-Antworten darf und muss der Testleiter die Testperson sogar dazu auffordern, noch mehr dazu zu sagen, wenn sie noch mehr dazu weiß (Petermann & Petermann, 2007).

Diese Vorgangsweise führt zu Problemen hinsichtlich der Interpretierbarkeit und inhaltlichen Eindeutigkeit dieser Skalen (im Sinne der Validität), hinsichtlich der Skalierung dieser Untertests und die praktische Handhabung (für Testleiter) betreffend, die nachfolgend beschrieben werden.<sup>10</sup> Da diese Probleme bei einer *dichotomen* Verrechnung nicht (oder deutlich geringer) aufträten, wird im Kapitel 5.11 diskutiert, ob eine solche dichotome Verrechnung zumindest ähnlich messgenau wäre wie die dreikategorielle und somit eine Alternative darstellen könnte.

---

<sup>10</sup> Kubinger (1983), sowie Steurer (1988) untersuchten und kritisierten diese und weitere damit zusammenhängende Probleme (wie mangelnde Auswertungsobjektivität, fragliche Eindimensionalität) bereits für den HAWIK bzw. den HAWIK-R, allerdings mit weitaus größeren Datensätzen und anderen, entsprechend präziseren statistischen und testtheoretischen Methoden.

## **4.2 Inhaltliche Kritikpunkte (Validität): die Fähigkeit, 2-Punkt-Antworten zu geben, als eigene Fähigkeitsdimension und mögliche Zusammenhänge mit leistungsbezogenen Persönlichkeitsvariablen.**

### 4.2.1 Klärung der Fragestellung

In den Testungen dieser Studie konnten viele Verhaltensbeobachtungen gemacht werden, die das „Suchen“ einer gültigen 2-Punkt-Antwort durch Testpersonen betraf, die auf eine Frage schon eine, möglicherweise auch mehrere 1-Punkt-Antworten gegeben hatten; es zeigten sich dabei interindividuelle Unterschiede, ob und wie viele alternative Antworten zusätzlich zur zuerst genannten geäußert wurden, und auch, wie diese weiteren Antworten zustande kamen: so gab es Testpersonen, die eher *andere Formulierungen* der inhaltlich bereits gegebenen Antwort suchten, und andere Testpersonen, die eher *inhaltlich andere* Antworten äußerten; diese eher unsystematisch wahrgenommenen Eindrücke der Testleiter verdichteten sich zu der Überlegung, ob es zusätzlich zu den zu messen intendierten Fähigkeitsdimensionen (Wortschatz, sprachbezogenes logisches Denken...) noch zumindest *eine* wesentliche Dimension gäbe, von der es abhängt, ob 2-Punkt-Antworten gegeben werden (können) oder nur 1-Punkt-Antworten. Dies könnte einerseits eine eigene Fähigkeitsdimension darstellen, andererseits auch von Persönlichkeitsvariablen abhängen.

Um dieses Problem zu erläutern, werden hier beispielhaft Fragen mit einigen dazugehörigen Antwortmöglichkeiten dargestellt:

### 4.2.2 Darstellung einiger fraglicher Antwortkodierungen des Untertests *Gemeinsamkeiten Finden*

Zum Item GF 10 „Was haben STIRNRUNZELN und LÄCHELN gemeinsam?“ gibt es (unter anderem) folgende angeführte Antwortmöglichkeiten:

Als 2-Punkt-Antworten gelten: „Gesichtsausdrücke; Ausdrücke, die man mit Gesicht/Mund/Lippen erzeugt“; „Mimik“; „Wie man mit Gesicht/Mund/Lippen seine Gefühle zeigt“ [...] „Gesten zur Beschreibung von Gefühlen/Emotionen/Stimmungen“ [...]

Als 1-Punkt-Antworten gelten unter anderem: „Gesten (N)“, „Ausdrucksbewegungen des Gesichtes (N)“; „Ausdrucksmöglichkeiten (N)“; „Mienen im Gesicht (N)“ (alle: Petermann & Petermann, 2007, S. 172)

Die hier aufgezählten 1-Punkt-Antworten sind mit einem (N) gekennzeichnet, was bedeutet, dass der Testleiter nachfragen muss, und zwar mit den Worten „Was genau meinst du damit?“ oder „Erzähle mir noch mehr darüber“, sodass die Testperson dazu angeregt wird, eine 2-Punkt-Antworten zu nennen (wenn sie sie weiß). Der Messintention entsprechend sollten die 1- und 2-Punkt-Antworten *nur* von der zu messenden Fähigkeit abhängen (also *eindimensional* messen, vgl. Kubinger, 2006). Das Manual nennt als zu messende Fähigkeiten „verbales Schlussfolgern und Konzeptbildung [...] außerdem auditives Verständnis, Gedächtnis, Unterscheidung zwischen unwichtigen und wichtigen Anteilen und verbalen Ausdruck“ (Petermann & Petermann, 2007, S. 38).

Bei genauer Betrachtung fällt aber auf, dass einige der Antwortmöglichkeiten, die unterschiedlich bewertet werden sollen, sich eher durch ihre *Formulierung* als durch den sprachlich ausgedrückten *logischen Zusammenhang* unterscheiden, z.B. „Mienen im Gesicht“ (1 Punkt) – „Mimik“ (2 Punkte) oder „Ausdrucksbewegungen des Gesichtes“ (1 Punkt) – „Ausdrücke, die man mit [dem] Gesicht erzeugt“ (2 Punkte). Angenommen ein Kind beantwortet die erstgenannte Frage nach den Gesichtsausdrücken mit „Ausdrucksbewegungen des Gesichtes“ (1 Punkt) und wird daraufhin vom Testleiter gefragt, was genau es damit meine, so gibt es - abgesehen von der Möglichkeit, dass es auch ganz andere Antwortmöglichkeiten äußern kann - einerseits die Möglichkeit, dass es daraufhin die mit 2 Punkten bewertete (eigentlich nur anders formulierte) Antwort „Ausdrücke, die man mit [dem] Gesicht erzeugt“ gibt, oder, dass es das nicht tut, beispielsweise, weil es sich denkt, dass es das ja ohnehin schon (nur eben anders formuliert) gesagt habe und sich nicht wiederholen wolle, oder weil es einfach nicht davon ausgeht, dass allein das *Wiederholen* der Antwort in einer anderen *Formulierung* hilfreich sein könnte. Nun stellt sich die Frage, ob davon auszugehen ist, dass seine Fähigkeiten zum *verbalen Schlussfolgern und Konzeptbilden* dafür ausschlaggebend sind, dass es die 2-Punkt-Antworten geben kann, oder ob sich darin nicht vielleicht auch eine andere Fähigkeit zeigt, wie es das oben beschriebene Beispiel nahelegt: beispielsweise könnte das die Fähigkeit sein, zu erkennen, was der Testleiter von ihm erwartet, obwohl es eigentlich anders instruiert wurde; oder auch eine Persönlichkeitsvariable, wie Ausdauer oder Ehrgeiz.

Zum Item GF 13 „Was haben DICHTER und MALER gemeinsam?“ gibt es unter anderem die Antwortmöglichkeiten „[beide sind] künstlerisch“ (1 Punkt, N) – dagegen:



„[beide sind] Künstler“ (2 Punkte) (Petermann & Petermann, 2007, S.175), die sich in Bezug auf die Aufgabe, nämlich „Gemeinsamkeiten zu finden“, de facto nicht unterscheiden. Wiederum stellt sich die Frage, ob ein Kind, das schon „künstlerisch“ gesagt hat, nach der Aufforderung des Testleiters (noch mehr dazu zu sagen), wirklich deswegen das Wort „Künstler“ erwähnt, weil es die Antwort (im Sinne der zu messenden Fähigkeit) *besser weiß* als ein Kind, das bei „künstlerisch“ bleibt (vielleicht wieder nur deshalb, weil es sich nicht wiederholen möchte).

Noch deutlichere Zweifel daran, ob die Fähigkeit zur 2-Punkt-Antwort ein schlichtes *Mehr* an der Fähigkeit zur 1-Punkt-Antwort ist, sind bei einigen 1-Punkt-Antworten angebracht, die nicht mit (N) gekennzeichnet sind, bei denen also auch nicht nachgefragt werden darf. So listet das Manual zur Frage nach den Gemeinsamkeiten von Dichter und Maler folgende Antworten auf: „[beide] verdienen Geld mit ihrer Kunst“ (1 Punkt) – dagegen: „schaffen Kunst“; „tun kreative Dinge“ (jeweils 2 Punkte) (alle: Petermann & Petermann, 2007, S. 175).

Ebenso ohne Nachfrage mit einem Punkt zu bewerten sind die zu „Was haben ERLAUBNIS und EINSCHRÄNKUNG gemeinsam?“ gehörigen Antwortmöglichkeiten: „Direktiven; Gebote; Normen“, während die fast gleichbedeutenden Worte „Regeln; Vorschriften; Freiheitsgrade“ mit zwei Punkten zu bewerten sind (Petermann & Petermann, 2007, S. 183). Da der Testleiter hier nicht nachfragen darf (weil erstere Antwortmöglichkeiten ohne „N“ angeführt sind), scheint es umso mehr von Persönlichkeitsfaktoren (oder situativen Aspekten, Aspekten der TI-Tp-Interaktion usw.) abzuhängen, ob ein Kind noch weitere (inhaltlich ganz ähnliche) Antworten äußert oder nicht.

#### 4.2.3 Darstellung einiger fraglicher Antwortkodierungen des Untertests *Wortschatz-Test*

Ähnliche Beispiele sind aus dem *Wortschatz-Test* (WT) zu nennen: Dieser „wurde entwickelt, um das Wortwissen eines Kindes und seine Begriffsbildung zu erfassen. Er misst zudem den kindlichen Wissensschatz, die Lernfähigkeit, das Langzeitgedächtnis und den Stand der Sprachentwicklung“ (Petermann & Petermann, 2007, S. 38). Wieder stellt sich beim Vergleich vieler Antworten die Frage, ob ein Kind, welches die jeweilige 2-Punkt-Antworten gibt, wirklich mehr der zu messenden Fähigkeiten demonstriert als

ein Kind, das die jeweils aufgezählten 1-Punkt-Antworten äußert. Zur Illustration werden nun einige besonders auffällige Antwortmöglichkeiten aufgezählt:

Zu WT 11: „Was bedeutet MUTIG?“:

„sich etwas zutrauen“ (1 Punkt) – „sich etwas trauen“ (2 Punkte);

„jemanden aus der Gefahr retten (N)“ (1 Punkt) – „sich für andere opfern“ (sic!; 2 Punkte)

Zu WT 13: „Was bedeutet UNSINN?“:

„alberne/dumme/verrückte Dinge tun“ (je 1 Punkt) – „albernes Zeug“ ; „Quatsch“; „Dummheit“ (je 2 Punkte).

Zum WT 24: „Was bedeutet PRÄZISE?“:

„genau (N)“; „richtig“, „auf den Punkt (N)“ (jeweils 1 Punkt) – „exakt“, „genaue/r Messung/Betrag“; „absolut richtig“ (jeweils 2 Punkte)

(alle: Petermann & Petermann, 2007, S. 215)

#### 4.2.4 Darstellung einiger fraglicher Antwortkodierungen des Untertests *Allgemeines Verständnis*

Der Untertest *Allgemeines Verständnis* (AV) „wurde entwickelt, um verbales Schlussfolgern und verbale Konzeptualisierung zu messen. Außerdem [...] sprachliches Verständnis, sprachlichen Ausdruck, [...] Wissen um konventionelle Verhaltensstandards, soziales Urteil und soziale Reife, sowie den gesunden Menschenverstand“ (Petermann & Petermann, 2007, S. 38).

Zu diesem Untertest gibt es ähnliche Beispiele wie bisher dargestellt, bei denen sich die 1- und 2-Punkt-Antworten stark ähneln oder sich auf einer möglicherweise anderen, als der hier zu messen intendierten Dimension unterscheiden. Zusätzlich dazu gilt aber beim Untertest *Allgemeines Verständnis* noch die Regel, dass „bei Aufgaben mit mehrere allgemeinen Konzepten [...] die Antwort des Kindes mindestens zwei *unterschiedliche* allgemeine Konzepte beinhalten [muss], um sie mit 2 Punkten bewerten zu können“ (Petermann & Petermann, 2007, S. 254).

Ein Beispiel dafür ist das Item AV 9: „Nenne mir einige Gründe, warum man das Licht ausschalten sollte, wenn man es nicht braucht.“ Eine Testperson, die schon eine richtige Antwort aus dem Konzept: „Wissen darüber, dass es wichtig ist, *sparsam mit Energie* oder anderen *Umweltressourcen* umzugehen“ gegeben hat, muss, um zwei Punkte zu

bekommen, auch passend zum anderen Konzept: „Wissen darüber, dass man dadurch *Geld* sparen kann“ antworten (alle Zitate: Petermann & Petermann, 2007, S. 266). Hat nun eine Testperson „um Strom zu sparen“ geantwortet, so ist diese Antwort nur für das erste Konzept gültig, obwohl bei den Testungen an der Test- und Beratungsstelle mehrmals der Eindruck entstanden ist, dass Testpersonen der Meinung sind, durch Verwendung des Wortes *sparen* damit auch schon *Geld zu sparen* (zweites Konzept) zum Ausdruck gebracht zu haben, was man aber nicht so bewerten darf. Wiederum ist plausibel, dass andere Eigenschaften als die hier zu messen intendierte Fähigkeit systematisch beeinflussen, ob Testpersonen noch eine weitere Antwort (wie „um Geld zu sparen“) äußern und damit zwei Punkte erhalten, oder nicht.

#### **4.3 Statistische Untersuchungen der Frage: „2-Punkt-Antworten als eigene Fähigkeitsdimension“**

Im Rahmen statistischer Untersuchungen wurde einerseits versucht, eine eigene Dimension „Fähigkeit zur 2-Punkt-Antwort“, die in allen drei Untertests zum Tragen kommen könnte, zu finden. Dies geschah mittels einer Faktorenanalyse. Andererseits wurde untersucht, ob die Häufigkeiten der 2-Punkt-Antworten in jedem einzelnen Untertest Zusammenhänge zu bestimmten Persönlichkeitsvariablen zeigen.

Um das rechnerisch tun zu können, wurden für die drei fraglichen Untertests dichotomisierte Scores berechnet, bei denen sowohl die gültigen 1-Punkt-Antworten als auch die 2-Punkt-Antworten gleichermaßen mit 1 Punkt kodiert wurden, ungültige Antworten dagegen mit null Punkten. Diese Variablen wurden *GF\_dichotom\_leicht*, *WT\_dichotom\_leicht* und *AV\_dichotom\_leicht* genannt. Ein Teil der zu messenden Fähigkeiten (nämlich der, den man braucht, um zumindest eine 1-Punkt-Antworten zu geben), wird in diesen Variablen ausgedrückt.

Weiters wurde für jede Testperson für jede der drei Skalen einerseits die absolute Häufigkeit der 2-Punkt-Antworten als auch der relative Anteil der 2-Punkt-Antworten an den bearbeiteten Items berechnet. In diese Variablen fließen also auch die zu messenden Fähigkeitsdimensionen ein und außerdem – so die Hypothese – eine zusätzliche Dimension. Diese Variablen wurden *ZweierGF*, *ZweierWT* und *ZweierAV* (für die absoluten Häufigkeiten) und *Zweier\_rel\_GF*, *Zweier\_rel\_WT* und *Zweier\_rel\_AV* (für die relativen Häufigkeiten der 2-Punktantworten an den bearbeiteten Items) genannt.

Für alle folgenden statistische Berechnungen gilt:  $n = 41$  (außer anders angegeben).

#### 4.3.1 Faktorenanalyse zur Bestimmung einer gemeinsamen Fähigkeit zur 2-Punkt-Antwort über alle Untertests des Index *Sprachverständnis*

Um einen möglicherweise bestehenden gemeinsamen Faktor aufzufinden, der bei allen drei Untertests (zusätzlich zu den Fähigkeiten) die Häufigkeit von 2-Punkt-Antworten beeinflusst, nicht aber die Häufigkeit der „dichotom\_leicht“ (mit einem Punkt) kodierten Antworten, wurden die oben genannten sechs Variablen einer Hauptkomponentenanalyse zugefügt. Nach dem Eigenwertkriterium  $>1$  ergab sich eine Ein-Faktorenlösung, die 74% der Varianz erklärt (s. Tabelle 4.1). Alle sechs Variablen laden in vergleichbarer Form auf diesen Faktor mit Faktorladungen zwischen 0,76 und 0,89. Es zeigt sich somit kein eigener gemeinsamer Faktor der 2-Punkt-Antwort-Variablen, der nicht auch für die „dichotomen“ Scores gilt. (Ein versuchsweise extrahierter zweiter Faktor, der nur 11 % der Varianz erklärt, zeigt lediglich eine Differenzierung der beiden *Allgemeines Verständnis*-Variablen von den übrigen, nicht aber einen Faktor, der nur die 2-Punkt-Antworten betrifft.)

**Tabelle 4.1:** Ladungsmatrix der drei dichotomisierten Scores und der relativen Anteile der 2-Punkt-Antworten an den bearbeiteten Items der drei Untertests des Index SV; Hauptkomponentenanalyse

	1	2
GF_dichotom_leicht	,883	-,061
WT_dichotom_leicht	,864	-,249
AV_dichotom_leicht	,876	<b>,307</b>
Zweier_rel_GF	,874	-,258
Zweier_rel_WT	,890	-,266
Zweier_rel_AV	,756	<b>,610</b>

#### 4.4 Zusammenhänge der Häufigkeit der 2-Punkt-Antworten mit Persönlichkeitsvariablen

Um Zusammenhänge der Häufigkeiten der 2-Punkt-Antworten zu Persönlichkeitsvariablen aufzuzeigen<sup>11</sup>, wurden partielle Korrelationen der relativen Anteile der 2-Punkt-Antworten an den bearbeiteten Items (die Variablen *Zweier\_rel\_GF*, *Zweier\_rel\_WT* und *Zweier\_rel\_AV*) zu bestimmten Persönlichkeitsvariablen berechnet. Kontrollvariablen dabei waren jeweils die „dichotomisierten“ Skalenscores (*GF\_dichotom\_leicht*, *WT\_dichotom\_leicht*, *AV\_dichotom\_leicht*).

Als mögliche Persönlichkeitsvariablen, die einen Einfluss auf das Antwortverhalten hinsichtlich der 2-Punkt-Antworten haben könnten, kamen Einschätzungen zur *Leistungsmotivation in der Testsituation*, zur *Ausdauer in der Testsituation*, zur *Frustrationstoleranz als Reaktion auf Misserfolgerlebnisse in der Testsituation* und *Selbstüberzeugung* in Frage. Diese Variablen wurden von den Testleitern per Ratingverfahren mit den Werten 0=unterdurchschnittlich, 1= durchschnittlich und 2 = überdurchschnittlich kodiert.

##### 4.4.1 Ergebnisse der partiellen Korrelationen:

Zur Abschätzung der Signifikanz nicht-parametrischer Korrelationen (die Voraussetzungen für eine Produkt-Moment-Korrelation nach Pearson waren aufgrund der rangskalierten Persönlichkeitsvariablen nicht gegeben) existiert eine Berechnungsvorschrift für partielle Rangkorrelationen nach Spearman. Demnach ist der partielle Rang-Korrelationskoeffizient in einen Fishers Z-Wert zu transformieren, der dann unter Einbezug der Stichprobengröße zufallskritisch zu bewerten ist: Die entsprechende Formel lautet:

$$u = Z * \sqrt{N - 4} \quad (\text{Bortz \& Lienert, 2003, S.260}).$$

Bei einem (in diesem Zusammenhang sinnvoll erscheinenden) Signifikanzniveau von  $\alpha = 0,05$  (bei zweiseitiger Testung) beträgt die kritische Schranke  $u = 1,96$ , wobei

---

<sup>11</sup> Dies stellt natürlich keine klassische *Hypothesenprüfung* dar, da die Hypothesen ihren Ursprung in den gleichen realen Gegebenheiten – das heißt Erfahrungen in den Testsituationen – haben wie die Daten, an denen sie „geprüft“ werden, und ist daher nur als statistische Untermauerung einer Exploration zu verstehen!

höhere Werte von  $u$  einem statistisch signifikantem Ergebnis entsprechen ( $H_0: \rho = 0$ ).  
Durch Einsetzen von  $u = 1,96$  und  $n = 41$  zu

$$1,96 = Z * \sqrt{41 - 4} \quad \text{und durch Umformung}$$

$$Z = \frac{1,96}{\sqrt{37}} = 0,322$$

wurde ein Z-Wert von 0,322 ermittelt, was laut Bortz und Lienert (2003) einem Korrelationskoeffizienten von  $\rho = 0,31$  entspricht. Das bedeutet, dass niedrigere Korrelationen sich nicht signifikant von null unterscheiden, gleich hohe oder höhere Korrelationen als signifikant größer als null anzusehen sind. Tabelle 4.2 gibt die ermittelten partiellen Rangkorrelationskoeffizienten wieder.

**Tabelle 4.2:** partielle Rang-Korrelationskoeffizienten (Kontrollvariable: dichotomisierte Scores „leicht“) der relativen Anteile der 2-Punkt-Antworten an den bearbeiteten Items mit ausgewählten Persönlichkeitsvariablen

partieller Rang-Korrelationskoeffizient	Zweier_rel_ GF	Zweier_rel_ WT	Zweier_rel_ AV
Leistungsmotivation in der Testsituation	-0,02	0,11	0,20
Ausdauer in der Testsituation	0,03	0,33	0,29
Frustrationstoleranz	0,20	0,25	-0,04
Selbstüberzeugung	0,00	-0,19	-0,01

Da es sich hierbei nur um eine *näherungsweise* Signifikanzprüfung handelt, dienten diese Wert eher als erster Richtwert zur Orientierung als als Prüfgröße. Korrelationen in entsprechender Höhe zeigten sich nur zur Variable *Ausdauer in der Testsituation* und zwar beim *Allgemeinem Verständnis* (Spearman<sub>partiell</sub>:  $\rho = 0,29$ ) und beim *Wortschatz-Test* (Spearman<sub>partiell</sub>:  $\rho = 0,33$ ). Der nächstkleinere Korrelationskoeffizient betrug nur mehr 0,25 (nämlich zwischen *Frustrationstoleranz* und *Allgemeines Verständnis*) und wurde deshalb nicht mehr weitergehend untersucht.

Um eine genauere Abschätzung betreffend der statistischen Signifikanz und vor allem der Größe des Effekts erhalten zu können, wurden zusätzlich Varianzanalysen berechnet.<sup>12</sup>

Es wurde für jeden Untertest eine univariate Varianzanalyse berechnet. Die abhängigen Variablen waren die jeweiligen (absoluten) Häufigkeiten der 2-Punkt-Antworten des Untertests, der Faktor *Ausdauer in der Testsituation* und der jeweilige dichotomisierte Score wurde als Kovariate eingegeben.<sup>13</sup>

Die *Ausdauer in der Testsituation* wurde bei 6 Testpersonen mit unterdurchschnittlich, bei 20 mit durchschnittlich und bei 15 mit überdurchschnittlich bewertet.

#### 4.4.2 Ergebnisse der Varianzanalysen

Beim *Gemeinsamkeiten finden* zeigte sich wie erwartet kein signifikanter Effekt der *Ausdauer* auf die Anzahl der 2-Punkt-Antworten.

Beim *Wortschatz-Test (WT)* zeigt sich zusätzlich zum Effekt des dichotomen Scores *WT\_dichotom\_leicht* von  $\text{Eta}^2 = 0,772$  ( $p < 0,001$ ) ein Effekt von *Ausdauer in der Testsituation* von  $\text{Eta}^2 = 0,281$  ( $p = 0,002$ ). Die (unter Einbezug der Kovariate) geschätzten Randmittel der Häufigkeit der 2-Punkt-Antworten liegen bei den *unterdurchschnittlich ausdauernden* im Mittel bei 18,2, bei den *durchschnittlich ausdauernden* bei 21,3 und bei den *überdurchschnittlich ausdauernden* bei 22,0 (siehe Abbildung 4.1). Die paarweisen Vergleiche (mit  $\alpha$ -Adjustierung nach Bonferroni) zeigen signifikante Unterschiede zwischen den *unterdurchschnittlichen ausdauernden* und den *durchschnittlichen* (Differenz im Mittel 3,1 Zwei-Punkt-Antworten,  $p = 0,008$ ) und den *überdurchschnittlichen* (Differenz im Mittel 3,9 Zwei-Punkt-Antworten;  $p = 0,002$ ). Diesen Differenzen entsprechen auf Wertpunkt-Ebene je nach Alter etwa 1 bis 2 Wertpunkte, also etwa  $1/3$  bis  $2/3$  der Standardabweichung, was zwar in der

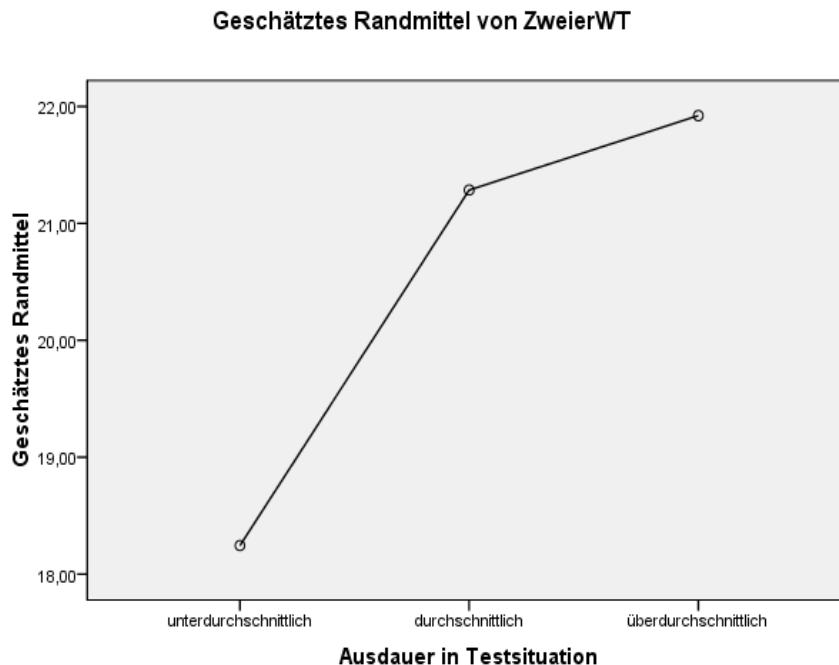
---

<sup>12</sup> Diese Vorgangsweise der Signifikanzprüfung soll aber wiederum keinesfalls den Eindruck erwecken, es handle sich hierbei um eine *Hypothesenprüfung*, da wiederum die Hypothese (nämlich, dass die *Ausdauer* einen Effekt auf die Häufigkeit der 2-Punkt-Antworten habe) auf den gleichen realen Gegebenheiten (Daten) basiert wie ihre statistische „Überprüfung“, die eben keine solche ist, sondern vielmehr eine Exploration.

<sup>13</sup> Es wurden hier die absoluten Häufigkeiten verwendet, da diese leichter zu interpretieren sind. Dass dies nicht zu verzerrten Ergebnissen führt, liegt daran, dass die Anzahl der bearbeiteten Items empirisch in keinem Zusammenhang zur *Ausdauer* stehen (was immerhin Verzerrungen hervorrufen könnte); daher korrelieren die Skalen der absoluten Häufigkeiten der 2-Punkt-Antworten auch mit den entsprechenden relativen Häufigkeiten nahezu perfekt mit  $r = 0,975$  (GF),  $r = 0,956$  (WT) und  $r = 0,968$  (AV).

Interpretation zu keinen gravierenden Unterschieden führen mag, aber dennoch zu beachten ist.<sup>14</sup>

**Abbildung 4.1:** *Wortschatz-Test*: Geschätzte Randmittel der univariaten Varianzanalyse der *Anzahl der 2-Punkt-Antworten*, Faktor: *Ausdauer in der Testsituation*, Kovariate: WT\_dichotom\_leicht



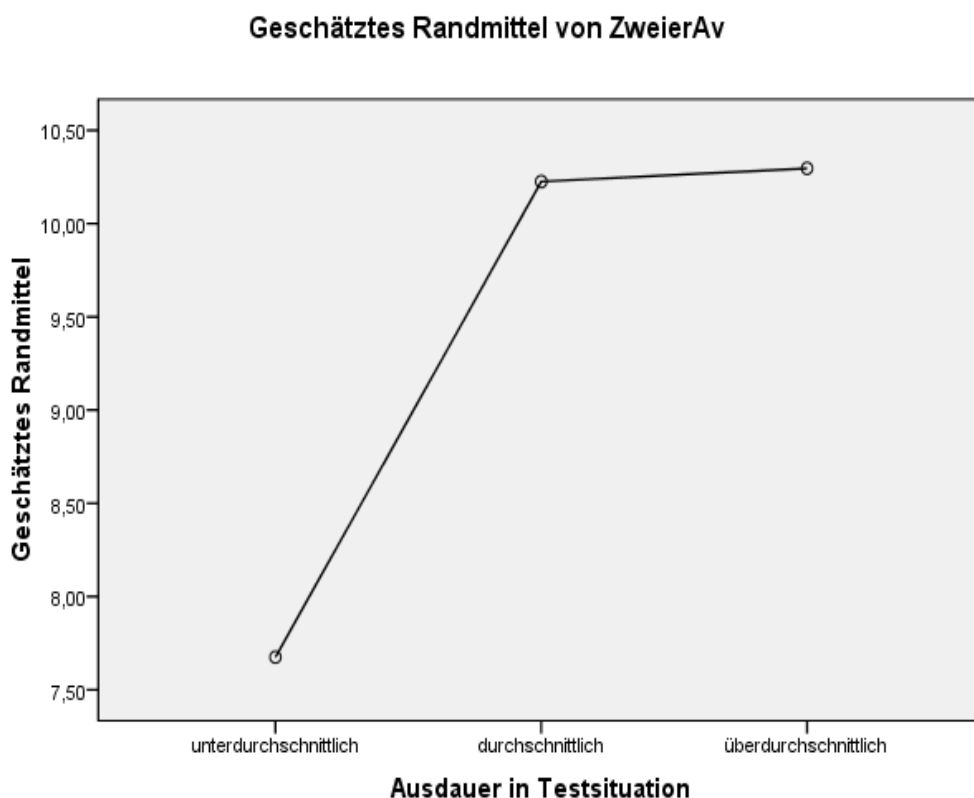
Beim *Allgemeinen Verständnis* zeigt sich ein Effekt des dichotomen Scores *AV\_dichotom\_leicht* von  $\text{Eta}^2 = 0,624$  ( $p < 0,001$ ) und ein Effekt von *Ausdauer in der Testsituation* von  $\text{Eta}^2 = 0,227$  ( $p = 0,022$ ); Die (unter Einbezug der Kovariate) geschätzten Randmittel der Häufigkeit der 2-Punkt-Antworten liegt bei den *unterdurchschnittlich ausdauernden* im Mittel bei 7,7, bei den *durchschnittlich ausdauernden* bei 10,2 und bei den *überdurchschnittlich ausdauernden* bei 10,3 (siehe Abbildung 4.2). Die paarweisen Vergleiche (mit  $\alpha$ -Adjustierung nach Bonferroni) zeigen signifikante Unterschiede zwischen den *unterdurchschnittlich ausdauernden* und den *durchschnittlichen* sowie den *überdurchschnittlichen* (Differenz im Mittel je 2,6 2-Punkt-Antworten,  $p = 0,012$  bzw.  $p = 0,014$ ). Diesen Differenzen entsprechen je nach

<sup>14</sup> (Tatsächlich unterscheiden sich die *unterdurchschnittlich ausdauernden* auf **Wertpunkt-Ebene** von den anderen beiden Gruppen mit etwa 3 Wertpunkten ( $p = 0,037$  bzw.  $p = 0,044$ , nach Bonferroni-Adjustierung für Mehrfachvergleiche), wobei dieser Unterschied natürlich auch einen Effekt von *Ausdauer* auf die 1-Punkt-Antworten beinhaltet und für die hier diskutierte Fragestellung nur bedingt aussagekräftig ist.)



Alter wiederum etwa 1 bis maximal 3 Wertpunkte, also etwa 1/3 bis maximal einer ganzen Standardabweichung, was durchaus beachtenswert ist.<sup>15</sup>

**Abbildung 4.2:** *Allgemeines Verständnis:* Geschätzte Randmittel der univariaten Varianzanalyse der Anzahl der 2-Punkt-Antworten, Faktor: *Ausdauer in der Testsituation*, Kovariate: *AV\_dichotom\_leicht*



#### 4.4.3 Auswirkungen auf Indexebene

Wieviele Wertpunkte (ein bis drei) es tatsächlich sind, die die oben dargestellten Unterschiede in der Anzahl der 2-Punkt-Antworten ergeben, ist nur von der Umrechnungstabelle der Rohwerte in Wertpunkte abhängig. So entsprechen beispielsweise beim *Allgemeinen Verständnis* einem Rohwert von 23-24 neun Wertpunkte, dem Rohwert von 25 zehn Wertpunkte und dem Rohwert von 26-27 elf Wertpunkte<sup>16</sup>, woraus ersichtlich wird, dass mit ein wenig „Pech“ die oben dargestellten Effekte von *Ausdauer in der Testsituation* in allen drei Untertests gemeinsam

<sup>15</sup> Tatsächlich unterscheiden sich die *unterdurchschnittlich ausdauernden* auf Wertpunkt-Ebene von den *durchschnittlich ausdauernden* mit etwa 2,9 (nach Bonferroni-Adjustierung nicht signifikant mit  $p=0,065$ ) und von den *überdurchschnittlichen* mit 4,3 ( $p=0,005$ ) Wertpunkten, wobei dieser Unterschied natürlich auch einen Effekt von *Ausdauer* auf die 1-Punkt-Antworten beinhaltet und für die hier diskutierte Fragestellung nur bedingt aussagekräftig ist.

<sup>16</sup> bei einem Kind von 12;8 Jahren bis 12;11 Jahren und 30 Tagen (vgl. Petermann & Petermann, 2007, S.370)

(Allgemeines Verständnis und Wortschatz-Test; Gemeinsamkeiten finden ist dabei ohne Belang) auch einen Unterschied von fünf Wertpunkten beinhalten können.

Daher wurde untersucht, ob sich der beschriebene Effekt von *Ausdauer in der Testsituation* auf die Untertests auch im *Index-Wert (Sprachverständnis)* empirisch zeigt. Zu diesem Zweck wurde eine Varianzanalyse mit dem *Index-Wert für Sprachverständnis* als abhängiger Variable und *Ausdauer in der Testsituation* als Faktor berechnet. Mögliche Unterschiede zwischen den Gruppen beinhalten in diesem Fall aber auch einen Effekt von *Ausdauer* auf die 1-Punkt-Antworten; die Ergebnisse sind daher für die hier diskutierte Fragestellung 2-Punkt-Antworten nur bedingt aussagekräftig (ein Einfluss der leistungsbezogenen Persönlichkeitsvariable *Ausdauer in der Testsituation* auf das Ergebnis in einem Leistungstest ist ja einigermaßen trivial). Jedenfalls aber würde ein auf Index-Ebene **nicht** feststellbarer Effekt von *Ausdauer* **gegen** die bisherige Argumentation sprechen.

Die Varianzanalyse zeigt einen signifikanten Effekt von *Ausdauer in der Testsituation* auf den Index-Wert für *Sprachverständnis* ( $\eta^2 = 0,993$ ,  $p < 0,001$ ). Der Indexwert der *unterdurchschnittlich ausdauernden* beträgt rund 114, der der *durchschnittlich ausdauernden* 132 und der der *überdurchschnittlichen* 134. Die Differenzen zwischen den *unterdurchschnittlich ausdauernden* und den anderen beiden Gruppen sind mit 18 bzw. 20 IQ-Punkten signifikant mit  $p = 0,006$  bzw.  $p = 0,003$  (nach Bonferroni-Adjustierung), was mehr als einer Standardabweichung entspricht. Dieser letzte Befund ist aber **nur** insofern bedeutsam, als er zumindest **nicht gegen** die weiter oben dargestellten Befunde in bezug auf die einzelnen Untertests (unter statistischer Kontrollen der Leistung in den dichotomisierten Scores) spricht.

#### **4.5 Zusammenfassung und Diskussion der Kritik an der dreikategoriellen Verrechnung hinsichtlich Validität**

In diesem Kapitel wurde diskutiert, inwiefern die Unterscheidung in 1- und 2-Punkt-Antworten bei den zum Index *Sprachverständnis* gehörigen Untertests – abgesehen von Problemen der Skalierung, die hier nicht diskutiert wurden – zu Problemen bezüglich ihrer Validität beitragen können. Es wurde vermutet, dass es zusätzlich zu den eigentlich zu messenden Fähigkeiten eine *eigene*, aber allen drei Untertests gemeinsame *Fähigkeitsdimension zur 2-Punkt-Antwort* gibt. Dies konnte mit einer zu diesem Zweck durchgeführten Faktorenanalyse nicht belegt werden. Darüber hinaus wurde untersucht,

inwiefern es Zusammenhänge der Häufigkeit von 2-Punkt-Antworten zu leistungsbezogenen Persönlichkeitsvariablen gibt. Hier zeigten sich moderate Zusammenhänge bei den Untertests *Wortschatz-Test* und *Allgemeines Verständnis* mit der Persönlichkeitsvariable *Ausdauer in der Testsituation*. Die Unterschiede der Häufigkeit der 2-Punkt-Antworten zwischen den Gruppen unterschiedlicher *Ausdauer* wurden durch Varianzanalysen bestätigt. Die Unterschiede betragen hinsichtlich ihrer potentiellen Auswirkung auf Ebene der Wertpunkte etwa 1/3 bis 2/3 (maximal 3/3) einer Standardabweichung.

Empirisch zeigen sich auf der Ebene der Wertpunkte dieser Untertests bzw. beim Index *Sprachverständnis* auch erhebliche Unterschiede von mehr als einer Standardabweichung zwischen *unterdurchschnittlich ausdauernden Testpersonen* einerseits und den *durchschnittlich oder überdurchschnittlich ausdauernden* andererseits. Letzterer Befund kann aber nicht eindeutig der hier diskutierten Verrechnung mit 1- und 2-Punkt-Antworten angelastet werden. Demensprechend gibt es auch keine eindeutigen Belege dafür, dass die Eignung des Indizes *Sprachverständnis* für eine entsprechende Interpretation im Rahmen einer Profilbetrachtung, wie sie beispielsweise vom *Wiener Modell zum Hochleistungspotenzial* verlangt wird, dadurch eingeschränkt werde. Ein besseres Abschneiden besonders ausdauernder Testpersonen allein aufgrund dieser leistungsbezogenen Persönlichkeitseigenschaft ist aufgrund der Ergebnisse auf Untertestebene zwar zu erwarten, kann hier jedoch nicht belegt werden.

#### **4.6 Kritik an der dreikategoriellen Verrechnung hinsichtlich Skalierung**

Ein weiterer problematischer Aspekt der dreikategoriellen Verrechnung der drei Untertests *Gemeinsamkeiten Finden*, *Allgemeines Verständnis* und *Wortschatz-Test* betrifft das Gütekriterium der Skalierung.

„Ein Test erfüllt das Gütekriterium *Skalierung*, wenn die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden,“ (Kubinger, 2006, S. 79)

Eine der Voraussetzungen für das Gütekriterium Skalierung stellen eindimensionale Messungen dar. Nicht-eindimensionale Messungen sind inhaltlich kaum interpretierbar oder nicht valide (vgl. Kubinger, 2006), wie im letzten Kapitel dargestellt wurde: darin konnte für die Untertests WT und AV ein Zusammenhang zwischen der Häufigkeit der 2-Punkt-Antworten und Persönlichkeitsvariablen festgestellt werden. Das würde

bedeuten, dass der Schritt zwischen „nicht-gelöst“ und einer 1-Punkt-Antwort wenigstens teilweise auf einer anderen *Dimension* liegen würde, als der zwischen einer 1-Punkt-Antwort und einer 2-Punkt-Antwort!

Doch selbst wenn die Ergebnisse, die nahelegen, dass dies zutrifft, angezweifelt werden, und außer Acht gelassen wird, dass die Autoren des HAWIK-IV auch alle Belege *für* die Eindimensionalität schuldig bleiben, so drängt sich doch der Eindruck auf, dass die besagten Schritte zumindest nicht *gleich groß* sind.<sup>17</sup>

Es besteht nämlich ein weiterer sehr naheliegender Grund, an der Güte der Skalierung des HAWIK-IV zu zweifeln: Es wird von den Testautoren nicht belegt, dass die unterschiedlichen Antwortkategorien, die mit 0, 1 und 2 bewertet werden, auch entsprechende Abstufungen der Fähigkeitsdimensionen repräsentieren; eine inhaltliche Betrachtung der Items legt im Gegensatz sogar die Vermutung nahe, dass dies nicht zutrifft; entweder weil der Schritt von „nicht-gelöst“ zur 1-Punkt-Antwort ein kleinerer ist als der Schritt zwischen der 1-Punkt-Antwort und der 2-Punkt-Antwort, oder weil der umgekehrte Fall zutrifft. Damit aber das Addieren der Punkte zu einem Gesamtscore zulässig ist, *müssen* die Differenzen der Punkte zwischen den unterschiedlichen Antwortkategorien den Unterschieden in der Fähigkeitsdimension entsprechen. Andernfalls würden die resultierenden Testwerte die empirischen Verhaltensrelationen eben *nicht* adäquat abbilden (vgl. die oben angeführte Definition von *Skalierung*)

Es liegen nicht genügend empirische Daten vor, um statistisch zu überprüfen, inwiefern die Voraussetzungen der Skalierung (z.B. die Eindimensionalität) verletzt wurden. Daher muss sich die Argumentation darauf beschränken, die Aufmerksamkeit des Lesers auf das Problem zu richten, und einige Beispiele vorzulegen:

Beispiele von Items, die die Voraussetzung der Skalierung in Zweifel ziehen lassen, wurden schon in den Kapiteln 3.1.2 und 3.1.3 angeführt; zwei davon sind aber hier noch einmal dargestellt (die hier angeführten Antwortmöglichkeiten sind nur ein kleiner Auszug der im Manual aufgezählten Antwortmöglichkeiten):

---

<sup>17</sup> Dies gilt sinngemäß auch für Untertest *Mosaik-Test*, der aufgrund der Konfundierung der Speed- und Powerkomponente offensichtlich nicht eindimensional misst; insbesondere fehlen auch Belege für die Skalierung der unterschiedlichen Lösungszeiten, wie dies schon für den HAWIK-III kritisiert wurde (vgl. Kubinger, 2006)

Das Item WT 11: „Was bedeutet MUTIG?“ hat unter anderem folgende Antwortkodierungen: „sich etwas zutrauen“ (1 Punkt) – „sich etwas trauen“ (2 Punkte); „jemanden aus der Gefahr retten (N)“ (1 Punkt) – „sich für andere opfern“ (2 Punkte) (Petermann & Petermann, 2007, S. 215).

Und zum Item GF 13 „Was haben DICHTER und MALER gemeinsam?“ gibt es unter anderem die Antwortmöglichkeiten „[beide sind] künstlerisch“ (1 Punkt, N) – dagegen: „[beide sind] Künstler“ (2 Punkte) (Petermann & Petermann, 2007, S. 175).

Das folgende (mit allen Antwortmöglichkeiten dargestellte) Item aus dem Untertest *Allgemeines Verständnis* zeigt einerseits, wie inhaltlich „nah“ Antworten aus den unterschiedlich zu bewertenden Kategorien sind, und andererseits, wie umfangreich die Aufzählung der möglichen Antworten ist, ohne dabei erschöpfend zu sein. (Die Reihenfolge wurde verändert, um ähnliche Antwortmöglichkeiten gegenüberzustellen.)

Item AV 16: „Warum soll man ein Versprechen einhalten?“

- „Menschen müssen sich gegenseitig vertrauen können“ (2 Punkte)
  - „damit andere einem trauen; es ist eine Frage des Vertrauens; zeigt dass man vertrauenswürdig ist; weil der andere einem vertraut hat“ (1 P.)
- „sie trauen einem und man möchte dieses Vertrauen haben“ (2 P.)
- „wenn einem die andere Person vertraut und wenn man die Freundschaft dieser Person schätzt, hält man sein Versprechen“ (2 P.)<sup>18</sup>
  - „um Freundschaften zu halten/bewahren; damit man seine Freunde nicht verliert“ (1 P.)
- „wenn jemand sein Versprechen nicht hält, kann ihm niemand glauben“ (2 P.)
  - „die anderen halten einen sonst für einen Lügner; die anderen glauben einem dann nicht mehr“ (1 P.)
  - „es zeigt, dass man auch meint, was man sagt/integer ist“ (1 P.)
  - „man hat sein Wort gegeben, und es wäre eine Lüge, dieses Wort zu brechen“ (1 P.)
- „die Vereinbarung zwischen zwei Menschen ist wie ein Vertrag und sollte respektiert werden; ein Versprechen ist ein sozialer Vertrag“ (2 P.)

---

<sup>18</sup> Anmerkung des Verfassers: man hält ein Versprechen nur *dann*, wenn man die Freundschaft dieser Person schätzt?

- „unser Sozialsystem beruht auf dem Glauben an/Vertrauen in Worte und Taten“ (2 P.)
  - „die anderen sind abhängig von einem/zählen/verlassen sich auf einen“ (1 P.)
  - „damit der andere nicht traurig/beleidigt/enttäuscht ist“ (1 P.)
  - „damit man die Gefühle anderer Menschen nicht verletzt“ (1 P.)
- „ein Versprechen ist ein Ehrenwort/eine Verpflichtung“ (2 P.)
  - „es ist eine Frage der Ehre“ (N, 1 P.)

(alle Zitate: Petermann & Petermann, 2007, S. 276)

Mit null Punkten sind *fehlende* Antworten, *falsche* Antworten und folgende *unzureichende* (bis falsche) Antworten zu bewerten:

- „es ist dein Wort (N)
- es ist moralisch/ethisch/ehrlich/höflich (N)
- es ist unfair, wenn man es nicht hält (N)
- andere wissen, welche Art von Mensch man ist (N)
- sonst ist der andere beleidigt (N)
- wenn man Versprechen nicht hält, können schlimme Dinge passieren/können andere Leute sauer werden
- wenn man was versprochen hat, muss man es auch einhalten; versprochen ist versprochen; man tut das Richtige; Versprechen sollten nicht gebrochen werden; es ist nicht nett, wenn man ein Versprechen bricht
- man würde lügen
- es ist ein Geheimnis; Geheimnisse darf man nicht preisgeben
- die Person wollte nicht, dass andere etwas davon erfahren; es könnte persönlich sein“

(Petermann & Petermann, 2007, S. 277)

Eine inhaltliche Überprüfung der Bewertungsvorgaben dieser Items – beispielsweise in dem Sinn, dass Experten versuchen, diese Antwortmöglichkeiten den unterschiedlichen Punkte-Werten zuzuordnen und dabei die vorgegebene Zuordnung zu replizieren – steht leider aus. So muss sich wohl jeder verantwortungsvolle Anwender des HAWIK-IV diese Frage selbst stellen.

#### **4.7 Kritik an der dreikategoriellen Verrechnung hinsichtlich der praktischen Handhabung**

Letztlich ist an der dreikategoriellen Verrechnung noch ein Kritikpunkt aus der Testpraxis anzuführen, der die praktische Handhabung betrifft: Ein wesentlicher Vorteil der Individualtestung gegenüber Gruppen- oder Computerverfahren stellen die vielfältigen Beobachtungsmöglichkeiten während der Interaktion und Leistungserbringung dar (vgl. Kubinger & Wurst, 2000, Kubinger, 2006). Die meisten der Items der dreikategoriell zu verrechnenden Untertests GF, WT und AV haben aber so viele Antwortmöglichkeiten (die sich darüber hinaus teilweise stark ähneln), dass die Aufmerksamkeit der Testleiter davon stark eingenommen werden kann. Zur Illustration wird hier auf die Auflistung der Antwortmöglichkeiten des Items 16 des Untertests *Allgemeines Verständnis* verwiesen (s.o.). Diese Beeinträchtigung der Beobachtungsmöglichkeiten aufgrund der Notwendigkeit, mehr als dreißig Antwortformulierungen zu lesen, um eine Übereinstimmung oder einen für die Punktebewertung wesentlichen Unterschied festzustellen, spricht jedenfalls stark gegen die praktische Nützlichkeit dieser Untertests. Dies gilt insbesondere auch für eine Diagnose nach dem *Wiener Diagnosemodell zum Hochleistungspotenzial*, zumal die Verhaltensbeobachtung einen wesentlichen Faktor der Erhebung leistungsbezogener Persönlichkeitsvariablen darstellt.

Sollte der testende Psychologe seine Aufmerksamkeit aber dennoch mehr auf die Interaktion und Beobachtung der Testperson lenken, so stellt die Ähnlichkeit vieler Antwortmöglichkeiten, die aber dennoch unterschiedlich bewertet werden sollen, die Verrechnungssicherheit bzw. Auswertungsobjektivität stark in Frage. Eine fehlende Eindeutigkeit bei der Bewertung analoger Antworten (mit ein oder zwei Punkten) kann aber zu deutlichen Testleitereffekten führen, wie Preusche (2007) für den HAWIK-III zeigen konnte.

Inwiefern diese Probleme durch eine dichotome Verrechnung dieser Untertests zu umgehen wären, wird im Rahmen des Kapitels „Reliabilität“ (unter 5.11) diskutiert.

## 5 RELIABILITÄT

„Die Reliabilität eines Tests beschreibt den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeitsmerkmal misst, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht“ (Kubinger, 2006, S. 45).

Die mittlere Reliabilität des HAWIK-IV für den Gesamt-IQ liegt laut Manual bei 0,97.

### 5.1 Split-half-Reliabilität

Die Messgenauigkeit (Reliabilität) eines Tests wird unter anderem daran geprüft, wie sehr seine einzelnen Teile dasselbe messen (innere Konstistenz); die bekannteste Methode dazu ist die „Split-half“-Methode: dabei werden die Items zwei Testhälften zugeordnet, meistens getrennt in Items mit gerader und ungerader Itemnummer; die Scores dieser Teile (in der Stichprobe) werden miteinander korreliert. Der erhaltene Korrelationskoeffizient wird dann zumeist (nach Spearman und Brown) mit der Formel  $r_{tt} = \frac{2r}{1+r}$  auf die ursprüngliche Länge des Testes aufgewertet. Mit Hilfe dieses Reliabilitätskoeffizienten ist es möglich, ein Konfidenzintervall für jeden Testwert zu berechnen, in dem der „wahre“ Testwert (der sich ergeben würde, hätte man einen messfehlerfreien, absolut exakten Test) bei einer vorher festgelegten Irrtumswahrscheinlichkeit (in der Testpraxis meist 5 %) liegt. Die Formel zur Berechnung dieses Konfidenzintervalls lautet:

$$T = x_v \pm z_{\alpha} \sqrt{s^2(1 - r_{tt})} \quad (\text{z.B. Kubinger, 2006}).$$

#### 5.1.1 Kritik an der Split-half-Reliabilität:

Diese Berechnungen der Messgenauigkeit und damit auch der Konfidenzintervalle sind, da es sich um korrelative Verfahren handelt, stark stichprobenabhängig und treffen daher nicht verlässlich für neue Testpersonen zu. Abgesehen von dieser Einschränkung zeigt sich noch ein prinzipielles Problem bei der Bestimmung der Messgenauigkeit mittels des Split-half-Reliabilitätskoeffizienten. Dies wird deutlich, wenn man in die Überlegungen mit einbezieht, dass diese Berechnungen ja implizit davon ausgehen, die einzelnen Items würden alleine deswegen korrelieren, weil sie das Gleiche oder zumindest Ähnliches messen, d.h., die gefundene gemeinsame Varianz wird letztlich als Beleg dafür interpretiert, dass hier ein und dasselbe Merkmal gemessen werde. Die Problematik dieses Schlusses kann anhand eines Gedankenexperiments dargestellt werden: Folgt man



nämlich diesem Gedanken, so wird klar, dass in Stichproben, wo viele der vorgegebenen Items *deutlich* zu schwierig oder *deutlich* zu leicht sind, die korrelativ erhobenen Reliabilitätskoeffizienten die Messgenauigkeit bei weitem überschätzen:

In diesem Gedankenexperiment wird eine Testskala bestehend aus 27 Items angenommen; die einzelnen Items sind dabei inhaltlich sehr inhomogen und prüfen unterschiedliche Fähigkeiten, die aber alle einigermaßen alterskorreliert bzw. anhängig von der bisherigen Beschulung sind, wie Rechnen, Wortschatz, Rechtschreibung.

Zur Stichprobe dieses Gedankenexperiments gehören zu gleichen Teilen Kinder mit etwa sechs Jahren, etwa elfjährige SchülerInnen und 16-jährige SchülerInnen. Die Items könnten dabei etwa so aussehen (s. Kasten 5.1):

**Kasten 5.1:** hypothetischer Test mit inhaltlich inhomogenen Items in drei globalen Schwierigkeitsgraden

Item 1: Wieviel ist  $1+1$ ?

Item 2: buchstabiere das Wort „Mama“

Item 3: was ist größer, ein Huhn oder ein Pferd?

Es folgen noch weitere 7 Items auf Niveau der ersten Klasse Volksschule.

Die nächsten sechs Items sind in der Schwierigkeit der ersten Klasse AHS angemessen (sie messen allerdings wieder keinesfalls das gleiche):

Item 11: Wieviel ist  $169 : 13$ ?

Item 12: Buchstabiere das Wort „Handyvertrag“.

Item 13: Welche Fische legen keine Eier?

Item 14: Was bedeutet die Abkürzung ÖBB?

Item 15: Nenne die ersten fünf Primzahlen.

Item 16: Nenne ein anderes (ähnliches) Wort für „Mediziner“.

Und dann folgen weitere zehn Items, die in ihrer Schwierigkeit etwa 16-jährigen angemessen sind:

Item 17: Jede gerade Zahl lässt sich als Summe zweier Primzahlen darstellen. Zeige das für 256.

Item 18: Wie hieß der letzte Papst?

Item 19: Wofür stehen die Buchstaben „DNA“?

Item 20: Was sind Winkelfunktionen?

...

Wenn wir nun diesen Test nach der gerade/ungerade-Methode halbieren, so ist nachvollziehbar, dass sehr kleine, etwa sechsjährige Kinder möglicherweise innerhalb der ersten zehn (2 mal 5) Items sehr unterschiedliche Ergebnisse erzielen (weil die Items ja zu ganz unterschiedlichen Fertigkeiten zählen), dass sie aber die Items ab Nummer 11 wahrscheinlich *alle* nicht lösen können, weil sie das (schlicht) in der Schule noch nicht gelernt haben. Folglich werden ihre beiden halben Testscores (bis auf seltene Ausnahmen) jeweils zwischen 0 und 5 streuen.

Für etwa 10- bis 12jährige Kinder wird es vermutlich eher so aussehen, dass *alle* Kinder die ersten zehn Items lösen können (weil sie ihnen deutlich zu einfach sind), die folgenden sechs Items dagegen in unterschiedlichem Maße, da sie in der Schwierigkeit etwa der Altersgruppe angepasst sind, aber eben unterschiedliche Fähigkeiten prüfen. Die letzten zehn Items werden sie höchstwahrscheinlich dagegen nicht lösen können. Ihre Testteile werden daher vermutlich v.a. zwischen den Werten 5 und 8 streuen.

Für 16-jährige wird es wiederum so aussehen, dass sie die ersten 16 Items (bis auf Ausnahmen) wahrscheinlich alle lösen können und nur die letzten zehn Items in unterschiedlichem Ausmaß; ihre Testteile streuen daher vor allem zwischen 8 und 13.

Das Ergebnis würde dann in etwa wie in Tabelle 5.1 aussehen (die Daten sind rein willkürlich im Rahmen der oben beschriebenen erwarteten Grenzen gewählt; es wurde innerhalb der Altersgruppen darauf geachtet, dass keine systematischen Zusammenhänge vorliegen, außer eben die zuvor beschriebenen, beispielsweise dass den meisten Sechsjährigen wohl alle Items ab Nummer 11 zu schwierig sind):

**Tabelle 5.1:** Scores der hypothetischen Testteile; innerhalb der im Text dargestellten Bedingungen zufällig gewählte Werte

Testperson	Testteil 1 (gerade)	Testteil 2 (ungerade)
Sechsjähriger 1	0	4
Sechsjähriger 2	1	3
Sechsjähriger 3	2	5
Elfjähriger 1	5	7
Elfjähriger 2	6	8
Elfjähriger 3	8	6
16jähriger 1	8	12
16jähriger 2	9	11
16jähriger 3	10	13

Der aus diesen Daten errechnete Wert der Split-half-Reliabilität beträgt (nach Aufwertung mit der Spearman-Brown-Formel)  $r_{tt} = 0,94$  (!), obwohl innerhalb der Altersgruppen z.T. negative, jedenfalls keine systematisch positiven Korrelationen vorliegen. D.h., diese Items messen im „passenden“ Schwierigkeitsbereich Unterschiedliches, dafür aber bei den meisten Items (je nach Alter) schlicht „das ist zu leicht“ (die Items haben also eine Lösungshäufigkeit nahe 1) oder: „das ist zu schwierig“ (haben also eine Lösungshäufigkeit nahe 0), was zu dieser hohen Korrelation der Testteile führt und den Eindruck erweckt, die Items würden die selbe Fähigkeit erfassen, was aber keinesfalls den Tatsachen entspricht.

Würde in diesem Gedankenexperiment den gleichen Personen aber nur die Items vorgegeben werden, die in ihrer Schwierigkeit halbwegs passen, und die anderen Items nicht, da die viel zu leichten und die viel zu schwierigen Items ohnehin keine Information bringen, so würde das ein deutlich realistischeres Bild der Messgenauigkeit ergeben: den Sechsjährigen würden demnach nur die ersten zehn Items, den Elfjährigen nur die mittleren sechs Items und den 16-jährigen nur die letzten zehn Items vorgegeben; dann würden (bei gleichem hypothetischen Antwortverhalten wie oben!) die Daten so aussehen (s. Tabelle 5.2):

**Tabelle 5.2:** Die Scores der hypothetischen Testteile aus Tabelle 5.1 nach Elimination der für jede Altersgruppe zu leichten und zu schwierigen Items

	Testteil 1 (gerade)	Testteil 2 (ungerade)
Sechsjähriger 1	0	4
Sechsjähriger 2	1	3
Sechsjähriger 3	2	5
Elfjähriger 1	0	2
Elfjähriger 2	1	3
Elfjähriger 3	3	1
16jähriger 1	0	4
16jähriger 2	1	3
16jähriger 3	2	5

Diesen Daten entspricht ein Reliabilitätskoeffizient (wieder aufgewertet nach Spearman und Brown) von  $r_{tt} = -0,21$ !

Der davor erhobene Reliabilitätskoeffizient von  $r_{tt} = 0,94$  wurde also zum Großteil dadurch verursacht, dass die eigentlich uninformativen, weil nämlich viel zu schwierigen oder viel zu leichten Items mitverrechnet wurden. Daraus lässt sich erkennen, dass der Reliabilitätskoeffizient nach der Split-half-Methode die wahre Messgenauigkeit weit überschätzen kann, wenn für die Testpersonen die Mehrzahl der Items entweder (deutlich) zu leicht oder zu schwierig sind, weil die den Items entsprechenden unterschiedlichen (!) Fähigkeiten (z.B. Rechnen, Rechtschreibung, Alltagswissen, Wortschatz) alle einigermaßen hoch mit dem Alter korrelieren, und daher überhaupt nur wenige Items ihrem Fähigkeitsniveau entsprechen (also eine Lösungshäufigkeit haben, die nicht nahe 0 oder nahe 1 ist). In diesem Gedankenexperiment wurden nur die Unterschiede in der Lösungswahrscheinlichkeit bedingt durch das *Alter* miteinbezogen, zusätzlich sind natürlich auch andere Faktoren wahrscheinlich, die dazu führen, dass für bestimmte Testpersonen viele Items klar zu leicht oder zu schwierig sind, ohne das gleiche zu messen (z.B. ein allgemeiner g-Faktor der Intelligenz).

Folgt man dieser Argumentation, so wird einsichtig, dass eine Betrachtung der Split-half-Reliabilitäten alleine kein adäquates Bild von der Messgenauigkeit einer Skala geben kann. Zusätzlich sind Angaben zu den relativen Lösungshäufigkeiten der Items vonnöten: sind diese nämlich für einen Großteil der Items nahe 0 oder 1, dann kann, wie oben gezeigt, auch eine relativ inhomogene Skala zu hohen Split-half-Reliabilitäten führen.

Offensichtlich ist auch, dass die Messgenauigkeit eines Untertests für die unterschiedlichen Niveaus (Schwierigkeitsgrade) erheblich variieren kann; das heißt, ein „quer“ über alle Schwierigkeitsgrade erhobener Split-half-Koeffizient stellt keine zuverlässige Schätzung der Messgenauigkeit für die *einzelnen* Fähigkeitsniveaus dar. (Zu beachten ist, dass aber eigentlich eben diese messgenaue Differenzierung *innerhalb* eines globalen Fähigkeitsniveaus von Interesse ist, und nicht, ob beispielsweise ein Test zuverlässig zwischen den Rechenfertigkeiten von Sechs- und Zehnjährigen differenzieren kann)!

## **5.2 Trennschärfe und Itemschwierigkeit**

Über die Split-half-Reliabilität hinausgehend sind auch Angaben zur Itemschwierigkeit und zur Trennschärfe einer Items zu beachten. Als Maß der Itemschwierigkeit gilt bei dichotomen Items die relative Lösungshäufigkeit; bei mehr als zwei-kategoriell

verrechneten Items wird die Itemschwierigkeit als Mittelwert der erreichten Punkteanzahl dividiert durch die Anzahl der maximal erreichbaren Punkte berechnet (vgl. Fisseni, 2004).<sup>19</sup>

Unter Trennschärfe versteht man die Korrelation eines Items mit dem Gesamtscore der Skala, zu dem das Item gehört, korrigiert um den Betrag, den das zu untersuchende Item zum Gesamtscore beiträgt (Lienert & Raatz, 1998).

Die Trennschärfeindizes sollen möglichst hohe Werte annehmen. Dabei ist es sinnvoll, wenn die Schwierigkeiten der Items gleichmäßig über den gesamten Bereich von 0,95 bis 0,05 verteilt sind (Kubinger, 2006). Ähnliches schlagen Lienert und Raatz (1998), zumindest im Falle mittlerer und hoher Trennschärfeindizes (um 0,6 oder höher) vor: hier sei es sinnvoll, die Verteilung der Schwierigkeitsindizes *breitgipfelig* um den Gipfel bei 0,5 anzulegen.

Im Falle *niedriger* Trennschärfe-Koeffizienten (etwa 0,3 bis 0,6) dagegen schlagen Lienert und Raatz (1998) vor, die Verteilung der Schwierigkeitsindizes eher *schmalgipfelig* mit dem Gipfel bei 0,5 zu gestalten. Dies bedeutet, dass wenn vor allem Items mit geringer Trennschärfe vorhanden sind, die Mehrzahl der Items eine relative Lösungshäufigkeit um 0,5 haben sollte – ein Vorgehen, dass so hohe „Schein-Reliabilitäten“ wie im oben dargestellten Gedankenexperiment weitgehend verhindern würde.

Im Folgenden werden daher nicht nur die im Manual des HAWIK-IV angegebenen Reliabilitätskoeffizienten dargestellt, sondern auch die entsprechenden (in der Stichprobe an der Test- und Beratungsstelle) empirisch gefundenen Split-half-Reliabilitätskoeffizienten der Untertests, und zusätzlich noch die Schwierigkeitsindizes und die Trennschärfekoeffizienten der Items, um so die Plausibilität dieser Reliabilitätsangaben abschätzen zu können.

---

<sup>19</sup> Diese Vorgangsweise ist streng genommen nur dann berechtigt, wenn die Bewertung der Antworten dem Gütekriterium der Skalierung entspricht, was im Falle der mehrkategorialen Untertests des HAWIK-IV angezweifelt werden kann. Dass diese Vorgangsweise hier dennoch gewählt wurde, kann einerseits damit argumentiert werden, dass sich die Verrechnung der Antworten im HAWIK-IV faktisch darauf stützt – alternative Interpretationsmöglichkeiten liegen gar nicht vor. Darüber hinaus liegt es daran, dass – den Vorgaberegeln des Tests folgend – die Häufigkeiten der Antworten im Zuge der Erhebungsphase dieser Studie gar nicht getrennt für 1- und 2-Punkt-antworten erhoben wurden: d.h., ob eine Testperson, die eine 2-Punkt-Anwort gab, auch die 1-Punkt-Anwort geben konnte, wurde nicht dokumentiert.

Die anhand der Stichprobe dieser Arbeit ermittelten Split-half-Reliabilitätskoeffizienten wurden zusätzlich zum üblichen Pearson-Korrelationskoeffizienten mittels des Rangkorrelationskoeffizienten nach Spearman berechnet, da teilweise die Voraussetzung der Normalverteilung verletzt war bzw. Ausreißer in den Daten vorlagen, was zu Verzerrungen des Pearson-Korrelationskoeffizienten führen kann.

Zur Berechnung der Reliabilitätskoeffizienten wurden nur diejenigen Items herangezogen, die auch wirklich bearbeitet wurden (also ohne die Items, die wegen Einstiegs- oder Abbruchregeln nicht vorgelegt wurden).

### **5.3 Mosaik-Test - Reliabilität, Itemtrennschärfe, Itemschwierigkeit**

#### **5.3.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe**

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,85$ . Die Reliabilitätskoeffizienten pro Lebensalter (in Jahren) liegen zwischen 0,80 und 0,89 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine Split-half-Reliabilität von  $r_{tt} = 0,76$  (Pearson) bzw.  $\rho_{tt} = 0,77$  (nach Spearman) errechnet.

#### **5.3.2 Itemtrennschärfe, Itemschwierigkeit**

Die ersten acht Items des *Mosaik-Tests* werden dichotom verrechnet (gelöst oder nicht gelöst); für die ersten beiden Items sind 0 oder 2 Punkte, für das dritte bis achte Item sind 0 oder 4 Punkte zu vergeben. Bei den Items 9 bis 14 werden zusätzlich Zeitgutpunkte verrechnet; bei diesen Items werden 0 Punkte für „nicht-gelöst“ und 4, 5, 6 oder 7 Punkte für „gelöst“ je nach Schnelligkeit verrechnet.

Die relativen Lösungshäufigkeiten (bzw. Itemschwierigkeiten) der dichotom zu verrechnenden ersten acht Items sind allesamt 1 oder nahe 1; die ersten sechs Items weisen eine Lösungshäufigkeit von 100%, das Item 7 von 95 % und das Item 8 von 98% auf; sie sind dementsprechend innerhalb dieser Stichprobe nahezu ohne Informationsgewinn. Da aber prinzipiell in einem Leistungstest genügend Aufgaben mit geringer Schwierigkeit enthalten sein müssen, um auch leistungsschwache Testpersonen erfassen zu können, ist dies auch sinnvoll (Lienert & Raatz, 1998). Die Trennschärfe-

Koeffizienten der ersten sechs Items sind mangels Varianz nicht zu berechnen, die der Items 7 und 8 betragen 0,26 und 0,25, was als sehr gering einzustufen ist.

Das Items 9 bis 14 wurden (unabhängig von der Schnelligkeit) in der vorliegenden Stichprobe von 100 % der Testpersonen (Item 9), von 95 % (Item 10), von 83 % (Item 11 und 12), von bei 63,4% (Item 13) und von immerhin 29,3% (Item 14) der Testpersonen gelöst (siehe Tabelle 5.3).

**Tabelle 5.3:** relative Lösungshäufigkeiten der Items 9 bis 14 des *Mosaik-Tests* (ohne Berücksichtigung der Zeit-Gut-Punkte)

Item-Nr.	Lösungshäufigkeit (Schwierigkeit)
9	1
10	0,95
11	0,83
12	0,83
13	0,63
14	0,29

Ob ein Item gelöst werden kann oder nicht (im Wesentlichen unabhängig von der Geschwindigkeit) stellt den *Power*-Teil dieses Untertests dar, zusätzlich werden bei den Items 9 bis 14 Zeitgutpunkte vergeben; diese repräsentieren damit die *Speed*-Komponente dieses Untertests.<sup>20</sup> Hinsichtlich des *Power*-Aspektes sind aufgrund ihrer relativen Lösungshäufigkeiten in der Stichprobe in erster Linie die Items 13 und 14 informativ, und mit Einschränkungen auch die Items 11 und 12; außerdem zeigt sich bezüglich des *Power*-Aspektes ein Deckeneffekt, da das zweitschwierigste Item von 63% der Testpersonen und das schwierigste Item noch immer von etwa 30 % gelöst wurde. Insgesamt weisen damit deutlich zu wenig Items eine mittlere oder niedrige relative Lösungshäufigkeit auf (Items 13 und 14); die *Power*-Komponente, die durch den *Mosaik-Test* erfasst werden sollte, wird also im oberen Leistungsbereich nur sehr unzureichend genau abgebildet.

---

<sup>20</sup> In der Verrechnung des Scores dieses Untertests in Wertpunkte werden diese beiden Anteile standardmäßig *nicht* getrennt verrechnet sondern summiert, allerdings existieren auch Tabellen für eine Auswertung *ohne* Zeitgutpunkte (Petermann & Petermann, 2007).

Die Verrechnung der Zeitgutpunkte führt hinsichtlich der Items 9 und 10, und in stärkerem Ausmaß hinsichtlich der Items 10, 11, 12, 13 und 14 zu einem deutlichen Informationszuwachs, dabei wird allerdings vor allem die *Speed*-Komponente erfasst. Die Itemschwierigkeitsindizes *inklusive Zeitgutpunkte* sind in der folgenden Tabelle (Tab. 5.4) angeführt.

**Tabelle 5.4:** Itemschwierigkeiten und Trennschärfen der Items 9 bis 14 des *Mosaik-Tests* (unter Berücksichtigung der Zeitgutpunkte)

Item-Nr.	Itemschwierigkeit (Mittelwert der Punkte / 7)	Trennschärfe (Spearman)	Trennschärfe (Pearson)
9	0,79	0,54	0,52
10	0,72	0,76	0,74
11	0,62	0,63	0,63
12	0,62	0,70	0,67
13	0,50	0,74	0,70
14	0,21	0,22	0,21

Abgesehen von der prinzipiell zu kritisierende Konfundierung von *Power*- und *Speed*komponente zeigen sich hier ausreichend hohe Trennschärfen. Abgesehen vom Bereich der schwierigsten Items zeigt sich eine relativ ausgewogene Verteilung der Itemschwierigkeiten. Der oben angeführte Deckeneffekt, der sich hinsichtlich der *Power*-Komponente zeigt, wird zwar durch die Verrechnung der *Speed*komponente weniger auffällig (so erreichte keine Testperson die Höchstpunkteanzahl beim Item 14) – dennoch ist der Bereich der schwierigeren Items (also mit Itemschwierigkeiten unter 0,5) nur durch ein einziges Item und damit deutlich zu schwach repräsentiert, was dazu führt, dass leistungsstarke Personen nicht ausreichend messgenau voneinander differenziert werden.

Da dennoch zumindest etwas weniger als die Hälfte der Items als informativ (im Sinne der Itemschwierigkeit) gelten kann, und die deutlich zu einfachen Items 1 bis 8 *bei annähernd allen Testpersonen* eine Itemschwierigkeit von 1 aufweisen, sind keine bedeutsamen „artifizialen“ statistischen Zusammenhänge und damit keine der oben beschriebenen groben Verzerrungen der Split-half-Reliabilitätsschätzungen innerhalb dieser Stichprobe zu erwarten: eine Berechnung der Reliabilität in drei etwa gleich großen Altersgruppen (Gruppe1: bis 9;11 Jahre, Gruppe 2: 10 bis 11;9 Jahre; Gruppe 3:



ab 12 Jahre) ergab dennoch z.T. deutlich niedrigere Reliabilitätskoeffizienten (nach Pearson bzw. Spearman, jeweils aufgewertet nach Spearman und Brown):

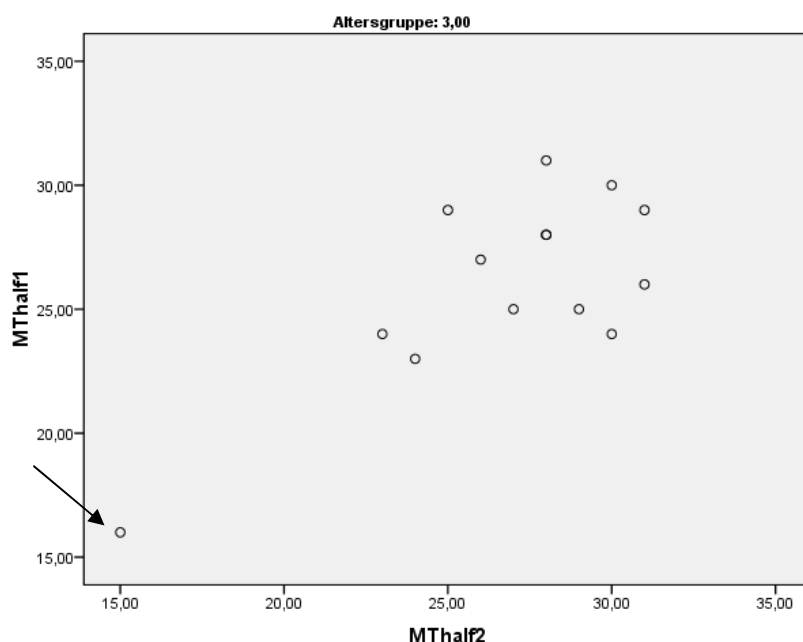
Altersgruppe 1:  $r_{tt1} = 0,59$  bzw.  $\rho_{tt1} = 0,54$

Altersgruppe 2:  $r_{tt2} = 0,46$  bzw.  $\rho_{tt2} = 0,46$

Altersgruppe 3:  $r_{tt3} = 0,86$  bzw.  $\rho_{tt3} = 0,63$

wobei der Pearson-Korrelationskoeffizient die Reliabilität in der Altersgruppe 3 überschätzt, da ein extremes Wertepaar vorliegt, welches den Koeffizienten verzerrt (siehe Abb. 5.1).

**Abbildung 5.1:** Streudiagramm der Scores der Testhälften des *Mosaik-Tests* in der Altersgruppe 3. Links unten das extreme Wertepaar, das zu einem überhöhten Pearson-Korrelationskoeffizienten führt.



### 5.3.3 Anmerkung zur Verringerung der Reliabilität aufgrund verringerter Varianz in den Teilstichproben

Da die der Reliabilitätsberechnung zugrundeliegenden Korrelationskoeffizienten alleine aufgrund einer niedrigeren Varianz in den Altersgruppen geringer ausfallen könnten (selbst wenn der Zusammenhang de facto in gleicher Art und Weise gegeben ist), könnte die verringerte Varianz in den Teilstichproben damit auch *rein rechnerisch* zu einer Verringerung der Reliabilitätskoeffizienten führen. Hunter und Schmidt (2004) schlagen vor, solche Verminderung der Korrelationskoeffizienten aufgrund verminderter Varianz anhand folgender Formel zu korrigieren:

$$R_{\text{kor}} = \frac{r \cdot \frac{s_g}{s}}{\sqrt{\left(\frac{s_g^2}{s^2} - 1\right) \cdot r^2 + 1}} \quad (\text{nach Hunter \& Schmidt, 2004, S. 199})$$

Der daraus resultierende „varianzkorrigierte“ Korrelationskoeffizient  $R_{\text{kor}}$  entspräche demnach der Korrelation innerhalb einer Altersgruppe, wenn in dieser die gleiche Varianz auftreten würde wie in der Gesamtstichprobe. Aus diesen korrigierten Korrelationskoeffizienten wurde dann (aufgewertet nach Spearman und Brown) die „varianzkorrigierten“ Reliabilitätskoeffizienten der Altersgruppen berechnet.

$s_g^2$  bezeichnet dabei die Varianz des Scores in der Gesamtstichprobe,  $s_i^2$  die Varianz des Scores in der jeweiligen Altersgruppe und  $R_{\text{kor},i}$  die varianzkorrigierte Korrelation in der jeweiligen Altersgruppe. Mit  $\text{rel}_{\text{kor},i}$  wurde der daraus berechnete varianzkorrigierte Split-half-Reliabilitätskoeffizient jeder Altersgruppe bezeichnet. Diese Vorgangsweise ist nur für Produkt-Moment-Korrelationen (nach Pearson) anwendbar (vgl. Hunter & Schmidt, 2004).

Folgende Werte wurden zur Berechnung herangezogen:

Varianz der Punkte in der Gesamtstichprobe:  $s_g^2 = 83,9$

Varianz in den Altersgruppen:  $s_1^2 = 62,3$ ;  $s_2^2 = 47,7$ ;  $s_3^2 = 56,0$

emp. Korrelationskoeffizienten i.d. Altersgruppen:  $r_1 = 0,416$ ;  $r_2 = 0,30$ ;  $r_3 = 0,748$ .

Da in der Altersgruppe 3 ein extremes Wertpaar vorlag – vgl. Abb. 5.1 –, das die Produkt-Moment-Korrelation (Pearson) nach oben verzerrte, wurde die Korrelation zusätzlich unter Ausschluss dieses Wertepaares nochmals berechnet, um zu einer realistischen Schätzung der Reliabilität in der Altersgruppe 3 zu kommen. Die entsprechenden Werte (nach Ausschluss des extremen Wertepaares) sind:

$$r_{3(-)} = 0,346, \quad s_{3(-)}^2 = 17,77, \quad s_{g(-)}^2 = 80,05$$

Daraus wurden die varianzkorrigierten Split-half-Reliabilitätskoeffizienten berechnet:

Altersgruppe 1, varianzkorrigiert:  $\text{rel}_{\text{kor}1} = 0,64$

Altersgruppe 2, varianzkorrigiert:  $\text{rel}_{\text{kor}2} = 0,56$

[Altersgruppe 3, varianzkorrigiert:  $\text{rel}_{\text{kor}3} = 0,90$ ]

Altersgruppe 3 (nach Ausschluss des Wertepaares), varianzkorrigiert:

$$\text{rel}_{\text{kor}3(-)} = 0,76$$

Es zeigt sich also, dass die *varianzkorrigierten* Reliabilitätskoeffizienten für die Altersgruppen in geringerem Maß von dem Reliabilitätskoeffizienten der Gesamtstichprobe von  $r_{tt} = 0,76$  abweichen, als es ohne Korrektur der Fall war. Da die Koeffizienten der Altersgruppen aber trotz Korrektur teilweise deutlich niedriger ausfallen, als es der Gesamtstichprobe entspräche, lässt sich erkennen, dass diese geringeren Reliabilitätskoeffizienten *nicht alleine* durch eine Varianzminderung erklärbar sind; vielmehr drücken sie aus, dass der *Mosaik-Test* innerhalb der Altersgruppen 1 und 2 weniger messgenau ist, als es der Split-half-Reliabilitätskoeffizient der Gesamtstichprobe erwarten lässt, und damit auch *zu wenig* messgenau für eine Interpretation als Einzelskala (bzw. im Sinne einer Profilinterpretation).

## **5.4 Bildkonzepte - Reliabilität, Trennschärfe, Itemschwierigkeit**

### **5.4.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe**

Die mittlere Reliabilität liegt laut Manual (über alle Altersgruppen von 6 – 16 Jahren) bei  $r_{tt} = 0,82$ , die Reliabilitätskoeffizienten pro Lebensalter (in Jahren) liegen zwischen 0,76 und 0,87 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine empirische Split-half-Reliabilität von  $r_{tt} = 0,72$  (nach Pearson) bzw. von  $\rho_{tt} = 0,70$  (nach Spearman) errechnet, was prinzipiell als zu niedrig für eine Interpretation als Einzelskala gelten kann: ein auf Basis dieser Reliabilität berechnetes Konfidenzintervall erstreckt sich nämlich (bei  $\alpha = 0,05$ ) über etwa 2 Standardabweichungen!

### **5.4.2 Itemtrennschärfe, Itemschwierigkeit**

Der Untertest *Bildkonzepte* beinhaltet 28 Items, die dichotom verrechnet werden, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt. In folgender Tabelle (5.5) werden einerseits die Anzahl der Personen, die diese Items bearbeiteten, als auch die relativen Lösungshäufigkeiten und die Trennschärfeindizes über 0,25 wiedergegeben:

**Tabelle 5.5:** Itemschwierigkeiten und Trennschärfen (insofern > 0,25) des Untertests *Bildkonzepte*

Item-Nr.	N	Itemschwierigkeit (rel. Lösungshäufigkeit)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
BK1	6	1		
BK2	6	1		
BK3	10	1		
BK4	10	1		
BK5	27	0,85	0,38	0,28
BK6	27	1		
BK7	41	1		
BK8	41	0,95		
BK9	41	0,98		
BK10	41	0,93		
BK11	41	0,93		
BK12	41	0,98	0,30	0,26
BK13	41	0,95		
BK14	41	1		
BK15	41	1		
BK16	41	0,88		
BK17	41	0,83		
BK18	41	0,63	0,27	
BK19	41	0,78	0,27	
BK20	41	0,95		
BK21	41	0,56		
BK22	40	0,50	0,45	0,45
BK23	40	0,53		
BK24	40	0,58		
BK25	40	0,18		
BK26	38	0,16	0,37	0,39
BK27	35	0,23		
BK28	32	0,03		

Die Items 1 bis 4 und 6 bis 15 und das Item 20 haben in der Stichprobe eine relative Lösungshäufigkeit von  $p > 0,9$ , was bedeutet, dass sie kaum Informationen hinsichtlich der Unterschiede der Leistungsniveaus der Testpersonen liefern, ihre Trennschärfeindizes sind auch dementsprechend niedrig.

Die relativen Lösungshäufigkeiten (Itemschwierigkeiten) der restlichen Items verteilen sich einigermaßen gleichmäßig auf das ganze Spektrum (zwischen 0 und 1), was prinzipiell als positiv einzuschätzen ist. Hinsichtlich der Trennschärfe zeigt sich allerdings, dass überhaupt nur zwei Items Trennschärfe-Koeffizienten größer als 0,3 aufweisen (nämlich BK 22 und BK 26 mit Trennschärfe-Koeffizienten von 0,45 und 0,39 bei Berechnung nach Spearman) und nur *ein* weiteres Item bei Berechnung der Korrelation nach Pearson (nämlich das Item BK 5). Darüber hinaus gibt es auch nur wenige weitere Items mit (ohnehin sehr geringen) Trennschärfeindizes von wenigstens 0,25: je nach Berechnung (nach Spearman oder Pearson) sind das die Items 5, 12, 18 und 19. Da also die deutlich zu leichten Items überwiegen und die in der Schwierigkeit angemessenen Items über relativ niedrige Trennschärfeindizes verfügen, ist durchaus zu erwarten, dass der (ohnehin nicht zufriedenstellende) Reliabilitätskoeffizient von  $r_{tt} = 0,72$  (bzw.  $\rho_{tt} = 0,70$ ) die Messgenauigkeit sogar noch überschätzt, dass sich also innerhalb der Altersgruppen noch geringere Split-half-Reliabilitätskoeffizienten finden lassen: eine Berechnung derselben in den drei Altersgruppen (Gruppe 1: bis 9;11 Jahre, Gruppe 2: 10 bis 11;9 Jahre; Gruppe 3: ab 12 Jahre) ergab demnach auch entsprechende Ergebnisse:

Altersgruppe 1:  $r_{tt1} = 0,88$  (Pearson) bzw.  $\rho_{tt1} = 0,90$  (Spearman)

Altersgruppe 2:  $r_{tt2} = 0,66$  (Pearson) bzw.  $\rho_{tt2} = 0,65$  (Spearman)

Altersgruppe 3:  $r_{tt3} = 0,35$  (Pearson) bzw.  $\rho_{tt3} = 0,22$  (Spearman).

Auch wenn diese Ergebnisse nur eingeschränkt über die Stichprobe hinaus zu generalisieren sind, so ergeben sich doch deutliche Anhaltspunkte dafür, dass die im Manual beschriebene Messgenauigkeit mit einem Split-half-Reliabilitätskoeffizienten von durchschnittlich  $r_{tt} = 0,82$  zumindest für die Stichprobe der eher überdurchschnittlich begabten Kinder und Jugendlichen ab 12 Jahren *nicht* gilt, sondern dass der Untertest *Bildkonzepte* in dieser Teilstichprobe deutlich zu ungenau misst!

#### 5.4.3 Anmerkung zur Verringerung der Reliabilität aufgrund verringerter Varianz in den Teilstichproben

Analog zur Anmerkung beim *Mosaik-Test* (Punkt 5.2.3) wurden auch für den Untertest *Bildkonzepte* varianzkorrigierte Reliabilitätskoeffizienten berechnet. Folgende Werte wurden dafür herangezogen:

Varianz der Punkte in der Gesamtstichprobe:  $s_g^2 = 7,35$

Varianz in den Altersgruppen:  $s_1^2 = 10,84$ ;  $s_2^2 = 5,14$ ;  $s_3^2 = 4,49$

Korrelationskoeffizienten i.d. Altersgruppen:  $r_1 = 0,788$ ;  $r_2 = 0,491$ ;  $r_3 = 0,21$

Die varianzkorrigierten Split-half-Reliabilitätskoeffizienten in den Altersgruppen betragen:

Altersgruppe 1: varianzkorrigiert:  $rel_{kor1} = 0,84$

Altersgruppe 2: varianzkorrigiert:  $rel_{kor2} = 0,72$

Altersgruppe 3: varianzkorrigiert:  $rel_{kor3} = 0,42$

Der varianzkorrigierte Reliabilitätskoeffizient der Altersgruppe 2 entspricht dem Reliabilitätskoeffizienten der Gesamtstichprobe von  $r_{tt} = 0,72$ , und auch der varianzkorrigierte Koeffizient der Altersgruppe 1 weicht in geringerem Maß davon ab, als es ohne Korrektur der Fall war. Ähnliches gilt für den Koeffizienten der Altersgruppe 3; dieser fällt aber trotz Korrektur deutlich niedriger aus, als es der Gesamtstichprobe entspräche. Es zeigt sich damit, dass die mangelnde Reliabilität dieses Untertests für die Jugendlichen über 12 Jahren *nicht* nur auf die verringerte Varianz zurückführbar ist, sondern dass die Messgenauigkeit der *Bildkonzepte* für diese Altersgruppe anhand der vorliegenden Daten tatsächlich als deutlich zu niedrig zu bewerten ist.

### 5.5 Matrizen-Test - Reliabilität, Trennschärfe, Itemschwierigkeit

#### 5.5.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,89$ , die Reliabilitäten pro Lebensalter (in Jahren) liegen zwischen 0,86 und 0,92 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine empirische Split-half-Reliabilität von  $r_{tt} = 0,78$  (nach Pearson) bzw. von  $\rho_{tt} = 0,80$  (nach Spearman) errechnet.

### 5.5.2 Itemtrennschärfe, Itemschwierigkeit

Der Untertest *Matrizen-Test* beinhaltet 28 Items, die dichotom verrechnet werden, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt; folgende Tabelle (5.6) gibt einerseits die Anzahl der Personen, die diese Items bearbeiteten, als auch die relativen Lösungshäufigkeiten und die Trennschärfeindizes über 0,25 wieder:

**Tabelle 5.6:** Itemschwierigkeiten und Trennschärfen (> 0,25) des Untertests *Matrizen-Test*

Item-Nr. <sup>21</sup>	N	Itemschwierigkeit (rel. Lösungshäufigkeit)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
MZ4	5	1		
MZ5	6	1		
MZ6	6	1		
MZ7	27	0,96		
MZ8	27	1		
MZ9	27	1		
MZ10	28	1		
MZ11	41	0,95	0,29	0,29
MZ12	41	0,98		
MZ13	41	0,98	0,28	0,26
MZ14	41	0,98		
MZ15	41	0,95		
MZ16	41	0,88	0,31	0,31
MZ17	41	1		
MZ18	41	0,88		
MZ19	41	0,95		
MZ20	41	0,98		
MZ21	41	0,98		

<sup>21</sup> Die Items 1, 2 und 3 wurden keiner Testperson vorgegeben und scheinen daher hier nicht auf.

MZ22	41	0,88	0,42	0,43
MZ23	41	0,76		
MZ24	41	0,59	0,29	0,31
MZ25	41	0,59	0,51	0,51
MZ26	40	0,40	0,40	0,41
MZ27	39	0,56		
MZ28	35	0,77		
MZ29	33	0,64	0,68	0,69
MZ30	31	0,52	0,54	0,53
MZ31	29	0,45		
MZ32	29	0,38	0,61	0,62
MZ33	27	0,26	0,40	0,39
MZ34	21	0,48	0,49	0,51
MZ35	17	0,18		

---

Die Items 1 bis 22 wurden in der Stichprobe der Test- und Beratungsstelle von zumindest 88 % der Testpersonen gelöst, was bedeutet, dass sie kaum Informationen hinsichtlich der Unterschiede der Leistungsniveaus der Testpersonen liefern, ihre Trennschärfen sind (bis auf das Item 22) auch dementsprechend niedrig.

Der Bereich der relativ leichten Items (also mit Lösungshäufigkeiten zwischen 0,8 bis 0,5) ist zwar mit sieben Items repräsentiert, allerdings weisen von diesen nur vier Items (die Items 24, 25, 39 und 30) Trennschärfekoeffizienten über 0,25 auf, was zu wenig erscheint, um in diesem Schwierigkeitsbereich Fähigkeitsdifferenzen der Testpersonen adäquat abbilden zu können.

Das gleiche gilt für den Bereich der schwierigeren Items: Items, die in der Stichprobe Lösungshäufigkeiten unter 0,5 und Trennschärfen über 0,25 aufweisen, gibt es überhaupt nur vier, wobei sich deren Lösungshäufigkeiten mit 0,48 (Item 34), 0,40 (Item 26), 0,38 (Item 32) und 0,26 (Item 33) schon auf eine leistungsstärkere Teilstichprobe beziehen, da das Item 32 beispielsweise nur mehr von 29 Tpn bearbeitet wurde, da der Test für die restlichen 12 Tpn aufgrund des Abbruchkriteriums schon abgebrochen worden war. Es wird deutlich, dass zu wenig aussagekräftige Items für den oberen und obersten Leistungsbereich vorhanden sind.



Eine Berechnung der Reliabilitätskoeffizienten in den drei Altersgruppen (Gruppe 1: bis 9;11 Jahre, Gruppe 2: 10 bis 11;9 Jahre; Gruppe 3: ab 12 Jahre) ergab folgende Ergebnisse:

Altergruppe 1:  $r_{tt1} = 0,84$  (Pearson) bzw.  $\rho_{tt1} = 0,79$  (Spearman);

Altersgruppe 2:  $r_{tt2} = 0,79$  (Pearson) bzw.  $\rho_{tt2} = 0,78$  (Spearman);

Altersgruppe 3:  $r_{tt3} = 0,75$  (Pearson) bzw.  $\rho_{tt3} = 0,78$  (Spearman).

Da die Ergebnisse der Teilstichproben denen der Gesamtstichprobe entsprechen, wurde auf die Darstellung der varianzkorrigierten Reliabilitätskoeffizienten (analog zu *Mosaik-Test* unter Punkt 5.2.3) verzichtet.

Der Matrizentest präsentiert sich damit als der messgenaueste Untertest des Index *Wahrnehmungsgebundenes logisches Denken*.

## **5.6 Gemeinsamkeiten Finden – Reliabilität, Trennschärfe, Itemschwierigkeit**

### 5.6.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,87$ , die Reliabilitäten pro Lebensalter (in Jahren) liegen zwischen 0,85 und 0,89 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurden empirische Split-half-Reliabilitätskoeffizienten von  $r_{tt} = 0,78$  (nach Pearson) bzw. von  $\rho_{tt} = 0,76$  (nach Spearman) errechnet.

### 5.6.2 Itemtrennschärfe, Itemschwierigkeit

Der Untertest *Gemeinsamkeiten Finden* beinhaltet 23 Items, wobei die ersten beiden dichotom, die restlichen 21 jedoch dreikategorial (mit 0, 1 oder 2 Punkten) verrechnet werden, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt. In Tabelle 5.7 sind einerseits die Anzahl der Tpn, die diese Items bearbeiteten, als auch die Trennschärfindizes über 0,25 und Itemschwierigkeiten angegeben:

**Tabelle 5.7:** Itemschwierigkeiten und Trennschärfen ( $> 0,25$ ) des Untertests *Gemeinsamkeiten Finden*

Item-Nr.	N	Itemschwierigkeit (erreichte Punkte / 2)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
GF1	5	1		
GF2	5	1		
GF3	26	0,96	0,29	
GF4	26	1,00		
GF5	41	0,95		
GF6	41	0,94	0,31	0,31
GF7	41	1,00		
GF8	41	0,84	0,58	0,55
GF9	41	0,84	0,26	0,31
GF10	41	0,77	0,68	0,67
GF11	41	0,77	0,51	0,47
GF12	41	0,92		
GF13	41	0,88	0,35	0,39
GF14	41	0,60	0,39	0,4
GF15	41	0,71	0,37	0,37
GF16	41	0,57	0,62	0,61
GF17	41	0,56	0,62	0,6
GF18	41	0,48		
GF19	40	0,38	0,25	0,29
GF20	40	0,41	0,35	0,36
GF21	39	0,42	0,47	0,48
GF22	38	0,28	0,49	0,49
GF23	38	0,21	0,29	0,27

Die Itemschwierigkeiten der Items verteilen sich relativ gut auf das Spektrum zwischen 0 und 1, wenn auch zu beanstanden ist, dass es zu wenig wirklich schwierige Items (als unter 0,3) gibt, was zu einer zu geringen Differenzierung im oberen Leistungsbereich führt. Darüber hinaus zeigt sich – der eher besser begabten Stichprobe entsprechend –,

dass sehr viele Items vorhanden sind, die von einem Großteil der Testpersonen gelöst werden: so werden die Items 1 - 9 von mehr als 80 % gelöst, weshalb diese Items innerhalb der vorliegenden Stichprobe nur als wenig informativ einzuschätzen sind. Ansonsten zeigen sich durchaus zufriedenstellende Trennschärfeindizes.

Die Berechnung der Reliabilitätskoeffizienten in den drei Altersgruppen (Gruppe 1: bis 9;11 Jahre, Gruppe 2: 10 bis 11;9 Jahre; Gruppe 3: ab 12 Jahre) ergab folgende Ergebnisse:

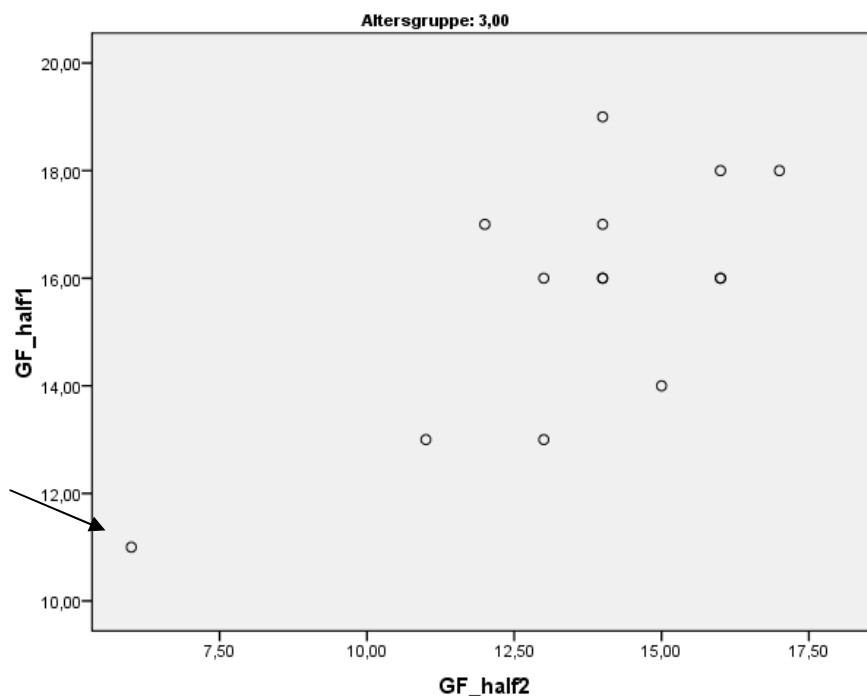
Altersgruppe 1:  $r_{tt1} = 0,77$  (Pearson) bzw.  $\rho_{tt1} = 0,77$  (Spearman);

Altersgruppe 2:  $r_{tt2} = 0,72$  (Pearson) bzw.  $\rho_{tt2} = 0,73$  (Spearman);

Altersgruppe 3:  $r_{tt3} = 0,83$  (Pearson) bzw.  $\rho_{tt3} = 0,69$  (Spearman),

wobei der letztgenannte Pearson-Koeffizient von 0,83 den Zusammenhang aufgrund eines Ausreißer-Wert-Paares überschätzen dürfte (siehe Abbildung 5.2).

**Abbildung 5.2:** Streudiagramm der Scores der Testhälften des Untertests *Gemeinsamkeiten Finden* in der Altersgruppe 3. Links unten das extreme Wertepaar, das zu einem überhöhten Pearson-Korrelationskoeffizienten führt.



Da die Ergebnisse der Teilstichproben ohnehin denen der Gesamtstichprobe entsprechen, wurde auf die Darstellung der varianzkorrigierten Reliabilitätskoeffizienten (analog zu *Mosaik-Test* unter Punkt 5.2.3) verzichtet.

## 5.7 Wortschatz-Test – Reliabilität, Trennschärfe, Itemschwierigkeit

### 5.7.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,90$ , die Reliabilitäten pro Lebensalter (in Jahren) liegen zwischen 0,84 und 0,92 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine empirische Split-half-Reliabilität von  $r_{tt} = 0,81$  (nach Pearson) bzw. von  $\rho_{tt} = 0,74$  (nach Spearman) errechnet, was wiederum unter den Angaben aus dem Manual liegt.

### 5.7.2 Itemtrennschärfe, Itemschwierigkeit

Der Untertest *Wortschatz-Test* beinhaltet 36 Items, wobei die ersten beiden Items mit 0 oder 1 Punkten bewertet werden. Die restlichen 34 Items jedoch werden dreikategorial (mit 0, 1 oder 2 Punkten) verrechnet, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt (so wurden die ersten beiden Items keiner Testperson vorgegeben). Tabelle 5.8 gibt einerseits die Anzahl der Tpn, die diese Items bearbeiteten, als auch die Trennschärfeindizes über 0,25 und Itemschwierigkeiten wieder:

**Tabelle 5.8:** Itemschwierigkeiten und Trennschärfen (> 0,25) des Untertests *Wortschatz-Test*

Item-Nr. <sup>22</sup>	N	Itemschwierigkeit (erreichte Punkte / 2)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
WT5	5	0,90	0,63	0,71
WT6	5	1		
WT7	27	1		
WT8	27	1		
WT9	41	0,99	0,31	0,26
WT10	41	0,99		
WT11	41	1		
WT12	41	0,98	0,26	
WT13	41	0,93	0,26	

<sup>22</sup> Die Items 1, 2, 3 und 4 wurden keiner Testperson vorgegeben und scheinen daher hier nicht auf.

WT14	41	0,93		
WT15	41	0,93	0,36	0,32
WT16	41	0,94	0,31	0,34
WT17	41	0,98	0,36	0,32
WT18	41	0,95	0,26	
WT19	41	0,83	0,66	0,59
WT20	41	0,92	0,30	0,32
WT21	41	0,78		
WT22	41	0,84	0,51	0,52
WT23	41	0,61	0,46	0,45
WT24	41	0,55	0,75	0,76
WT25	41	0,65	0,62	0,56
WT26	41	0,70	0,65	0,59
WT27	41	0,84	0,39	0,40
WT28	40	0,66	0,64	0,57
WT29	40	0,69	0,59	0,48
WT30	40	0,56	0,64	0,67
WT31	40	0,85		
WT32	39	0,41	0,45	0,51
WT33	39	0,71	0,57	0,43
WT34	39	0,46	0,37	0,34
WT35	39	0,17	0,41	0,47
WT36	39	0,60	0,58	0,65

---

Die Itemschwierigkeiten der Items verteilen sich relativ gut auf das Spektrum zwischen 0,5 und 1. Für den Bereich unter 0,5 ist - stärker noch als beim Untertest *Gemeinsamkeiten Finden* - zu beanstanden, dass es deutlich zu wenig mittelschwierige bis schwierige Items (also mit einer Schwierigkeit unter 0,5) und vor allem nur ein einziges Item mit einer Itemschwierigkeit unter 0,4 gibt, was zu einer deutlich zu geringen Differenzierungsfähigkeit dieses Untertests im oberen Leistungsbereich führt. Darüber hinaus zeigt sich – der eher überdurchschnittlich begabten Stichprobe entsprechend – , dass sehr viele Items vorhanden sind, die von einem Großteil der

Testpersonen gelöst wurden: so wurden 15 Items von mehr als 90 % der Testpersonen gelöst, die diese Items bearbeiteten, weshalb diese Items innerhalb der vorliegenden Stichprobe nur als wenig informativ einzuschätzen sind.

Die Trennschärfeindizes der Items im mittleren Schwierigkeitsbereich zeigen sich mit Werten zwischen 0,45 und 0,76 als durchaus zufriedenstellend.

Da dennoch ein großer Teil der Items als deutlich zu leicht einzuschätzen ist, ist der Anteil der Items, die die relevante Eigenschaft (nämlich „Wortschatz“) wirklich messen, gegenüber den Items, die möglicherweise nur den Umstand, dass sie (je Altersgruppe) „ganz allgemein sprachlich zu leicht“ sind, abbilden, so gering, dass es zu unrealistisch hohen Reliabilitätskoeffizienten für die Gesamtstichprobe kommen könnte. In diesem Fall würde die (ohnehin unbefriedigende) empirisch gefundene Reliabilität von  $r_{tt} = 0,75$  (nach Pearson) bzw. von  $\rho_{tt} = 0,71$  (nach Spearman) für die Gesamtstichprobe die Messgenauigkeit noch immer überschätzen. Die Reliabilitätskoeffizienten innerhalb der einzelnen Altersgruppen müssten dann deutlich niedriger (und damit realistischer) ausfallen.

Die Berechnung der Reliabilitätskoeffizienten in den drei Altersgruppen (Gruppe 1: bis 9;11 Jahre, Gruppe 2: 10 bis 11;9 Jahre; Gruppe 3: ab 12 Jahre) ergab folgende Ergebnisse:

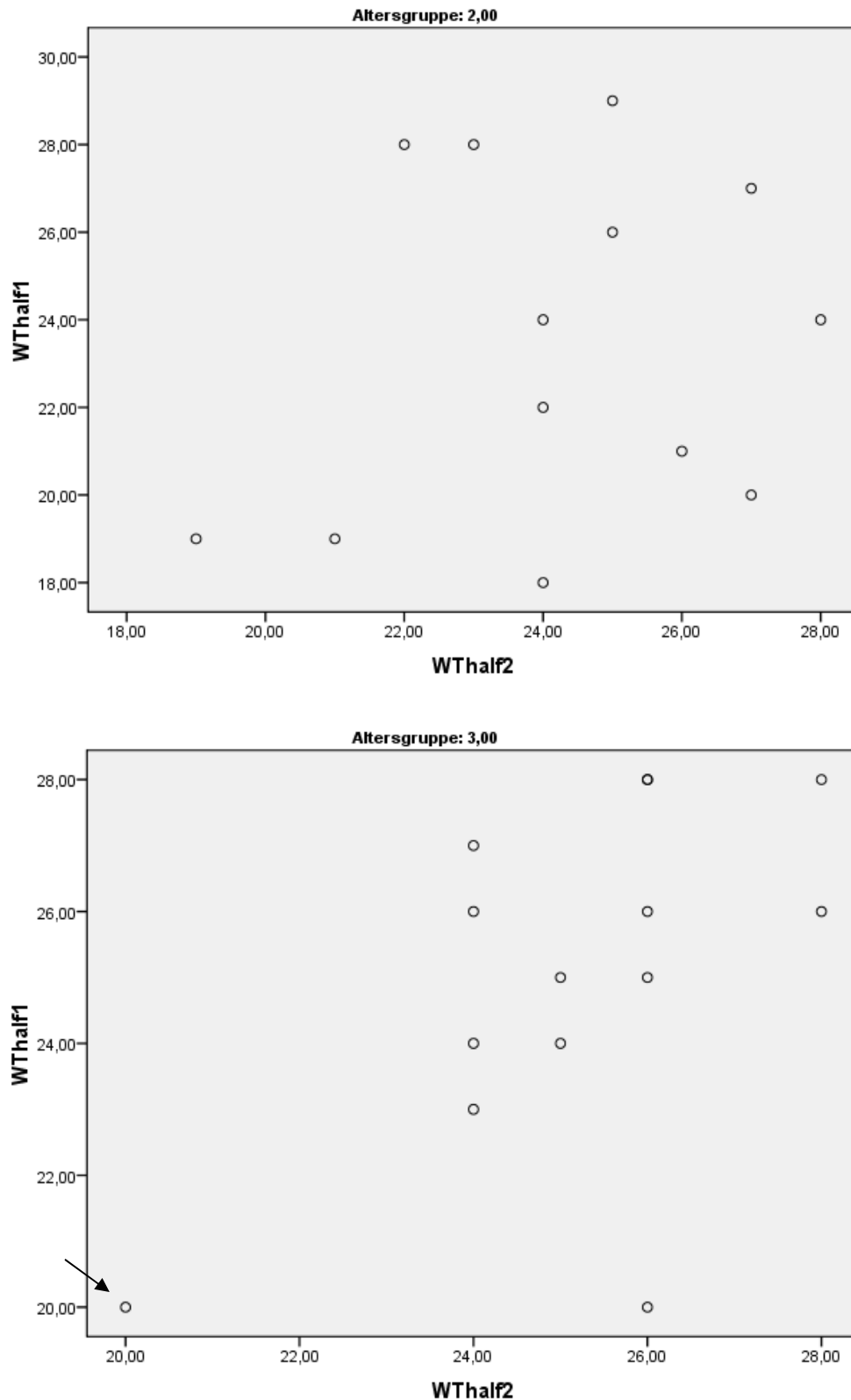
Altersgruppe 1:  $r_{tt1} = 0,77$  (Pearson) bzw.  $\rho_{tt1} = 0,78$  (Spearman);

Altersgruppe 2:  $r_{tt2} = 0,38$  (Pearson) bzw.  $\rho_{tt2} = 0,31$  (Spearman);

Altersgruppe 3:  $r_{tt3} = 0,73$  (Pearson) bzw.  $\rho_{tt3} = 0,67$  (Spearman),

wobei der letztgenannte Pearson-Reliabilitäts-Koeffizient von  $r_{tt3} = 0,73$  den wahren Zusammenhang aufgrund eines Ausreißer-Wert-Paares überschätzen dürfte (siehe Abb. 5.3).

**Abbildung 5.3:** Streudiagramme der Testhälften des Wortschatz-Tests in den Altersgruppen 2 und 3. Zu sehen ist der kaum vorhandene lineare Zusammenhang im oberen Streudiagramm (Altersgruppe 2) bei durchaus gegebener Varianz. Im unteren Streudiagramm (Altersgruppe 3) wird sichtbar, dass der lineare Zusammenhang durch den Pearson-Koeffizienten wegen des Ausreißer-Wertepaars (Pfeil) deutlich überschätzt wird.



Wie gering und daher wenig zuverlässig die Reliabilität des *Wortschatztests* ist, zeigt sich nicht nur in der absolut niedrigen Höhe des Split-half-Koeffizienten, insbesondere in Altersgruppe 2, sondern lässt sich auch dadurch illustrieren, dass der Korrelationskoeffizient zwischen den beiden Teilen des Wortschatztests (vor der Aufwertung nach Spearman und Brown) mit  $\rho = 0,59$  nicht bedeutend größer ist (in einem Fall sogar kleiner), als die Korrelationskoeffizienten zu den anderen Untertesthälften des Index *Sprachverständnis* (siehe Tabelle 5.9).

**Tabelle 5.9:** Korrelationskoeffizienten (Spearman) der Testhälften des *Wortschatz-Tests* zueinander und zu Testhälften anderer Untertests des Indizes *Sprachverständnis*

	WT_half1	WT_half2	AV_half1	AV_half2	GF_half1	GF_half2
WT_half1		0,59	0,55	0,46	0,77	0,49
WT_half2	0,59		0,51	0,49	0,46	0,41

### 5.7.3 Anmerkung zur Verringerung der Reliabilität aufgrund verringerter Varianz

Analog zum *Mosaik-Test* (Punkt 5.2.3) wurden auch für den Untertest *Wortschatz-Test* varianzkorrigierte Reliabilitätskoeffizienten berechnet. Folgende Werte wurden dafür herangezogen:

Varianz der Punkte in der Gesamtstichprobe:  $s_g^2 = 56,6$

Varianz in den Altersgruppen:  $s_1^2 = 59,4$ ;  $s_2^2 = 26,2$ ;  $s_3^2 = 16,9$

Korrelationskoeffizienten i. d. Altersgruppen:  $r_1 = 0,623$ ;  $r_2 = 0,232$ ;  $r_3 = 0,572$

Da in der Altersgruppe 3 ein extremes Wertpaar vorlag – vgl. Abb. 5.3 –, das die Produkt-Moment-Korrelation (Pearson) nach oben verzerren kann, wurde die Korrelation zusätzlich unter Ausschluss dieses Wertepaares nochmals berechnet, um zu einer realistischen Schätzung der Reliabilität in der Altersgruppe 3 zu kommen. Die entsprechenden Werte (nach Ausschluss des extremen Wertepaares) sind:

$$r_{3(-)} = 0,295, \quad s_{3(-)}^2 = 9,077, \quad s_{g(-)}^2 = 57,19$$

Die varianzkorrigierten Split-half-Reliabilitätskoeffizienten in den Altersgruppen betragen:



Altersgruppe 1, varianzkorrigiert:  $rel_{kor1} = 0,76$

Altersgruppe 2, varianzkorrigiert:  $rel_{kor2} = 0,50$

[Altersgruppe 3, varianzkorrigiert:  $rel_{kor3} = 0,88$ ]

Altersgruppe 3 (nach Ausschluss eines Wertepaares), varianzkorrigiert:

$$rel_{kor3(-)} = 0,76$$

Es zeigt sich also auch nach der Korrektur, dass die mangelnde Reliabilität dieses Untertests in der Altersgruppe 2 *nicht* nur auf die verringerte Varianz zurückführbar ist, sondern dass der Zusammenhang zwischen den Testteilen tatsächlich sehr gering ist (siehe auch Abbildung 5.3), und damit die Reliabilität zumindest für diese Teilstichprobe deutlich zu niedrig ausfällt. Für die Altersgruppe 3 zeigt sich aber nach der Korrektur ein der Gesamtstichprobe ( $r_{tt} = 0,75$ ) entsprechendes Ergebnis von  $rel_{kor3(-)} = 0,76$ .

## **5.8 Allgemeines Verständnis – Reliabilität, Trennschärfe, Itemschwierigkeit**

### **5.8.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe**

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,81$ , die Reliabilitäten pro Lebensalter (in Jahren) liegen bei den siebenjährigen Kindern bei nur  $r_{tt} = 0,71$ , in den restlichen Altersgruppen zumindest zwischen 0,77 und 0,87 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine deutlich höhere empirische Split-half-Reliabilität von  $r_{tt} = 0,89$  (nach Pearson) bzw. von  $\rho_{tt} = 0,88$  (nach Spearman) errechnet.

### **5.8.2 Itemtrennschärfe, Itemschwierigkeit**

Der Untertest *Allgemeines Verständnis* beinhaltet 21 Items, die dreikategoriell (mit 0, 1 oder 2 Punkten) verrechnet werden, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt; folgende Tabelle (5.10) gibt einerseits die Anzahl der Tpn, die diese Items bearbeiteten, als auch die Trennschärfeindizes über 0,25 und die Itemschwierigkeiten (berechnet als Mittelwert der erreichten Punkteanzahl dividiert durch die Anzahl der maximal erreichbaren Punkte, vgl. Fisseni, 2004) wieder:

**Tabelle 5.10:** Itemschwierigkeiten und Trennschärfen (> 0,25) des Untertests *Allgemeines Verständnis*

Item-Nr.	N	Itemschwierigkeit (erreichte Punkte /2)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
AV1	4	1		
AV2	4	1		
AV3	26	0,98		0,27
AV4	26	0,92	0,29	0,34
AV5	41	0,99		
AV6	41	1		
AV7	41	0,91	0,32	0,37
AV8	41	0,82	0,54	0,45
AV9	41	0,79	0,33	0,27
AV10	41	0,8	0,27	0,31
AV11	41	0,61	0,46	0,54
AV12	41	0,84		
AV13	41	0,35	0,58	0,63
AV14	41	0,5	0,61	0,73
AV15	41	0,66	0,67	0,67
AV16	39	0,53	0,42	0,40
AV17	38	0,67	0,59	0,58
AV18	38	0,28	0,57	0,64
AV19	37	0,31	0,59	0,69
AV20	34	0,32	0,65	0,65
AV21	30	0,27	0,63	0,71

Der Bereich der Itemschwierigkeiten zwischen 0,5 und 1 ist relativ gut durch Items abgedeckt, auch wenn die zu leichten Items (mit Schwierigkeiten über 0,9) überrepräsentiert sind. So zeigen acht Items Schwierigkeitsindizes von 0,5 bis 0,9, von diesen weisen fünf Items Trennschärfen zwischen 0,4 und 0,67 auf und zwei weitere Items Trennschärfen um 0,3; nur ein Item weist eine Trennschärfe unter 0,25 auf.

Auch der Bereich der eher schwierigen Items ist relativ gut vertreten: Itemschwierigkeiten unter 0,5 weisen immerhin fünf Items mit relativ hohen Trennschärfen (zwischen 0,57 und 0,65) auf; keines der Items weist Schwierigkeitsindizes unter 0,27 auf, was zu einer mangelhaften Differenzierung im oberen Leistungsbereich führen könnte.

Die Berechnung der Reliabilitätskoeffizienten in den drei Altersgruppen (Gruppe 1: bis 9;11 Jahre, Gruppe 2: 10 bis 11;9 Jahre; Gruppe 3: ab 12 Jahre) ergab folgende, weitgehend der Gesamtstichprobe entsprechende Ergebnisse:

Altersgruppe 1:  $r_{tt1} = 0,88$  (Pearson) bzw.  $\rho_{tt1} = 0,78$  (Spearman);

Altersgruppe 2:  $r_{tt2} = 0,85$  (Pearson) bzw.  $\rho_{tt2} = 0,86$  (Spearman);

Altersgruppe 3:  $r_{tt3} = 0,95$  (Pearson) bzw.  $\rho_{tt3} = 0,92$  (Spearman).

Anmerkung: Da die Ergebnisse der Teilstichproben ohnehin denen der Gesamtstichprobe entsprechen, wurde auf die Darstellung der varianzkorrigierten Reliabilitätskoeffizienten (analog zu *Mosaik-Test* unter Punkt 5.2.3) verzichtet.

## **5.9 Rechnerisches Denken – Reliabilität, Trennschärfe, Itemschwierigkeit**

### **5.9.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe**

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,89$ , die Reliabilitäten pro Lebensalter (in Jahren) liegen zwischen 0,83 und 0,93 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine empirische Split-half-Reliabilität von  $r_{tt} = 0,68$  (nach Pearson) bzw. von  $\rho_{tt} = 0,61$  (nach Spearman) errechnet. Ergänzend dazu muss aber festgestellt werden, dass nur 28 Testpersonen diesen Untertest bearbeiteten; der Grund dafür lag darin, dass bei den anderen 13 Testpersonen die vorhergehenden Untertests oder anderen psychologisch-diagnostischen Verfahren soviel Zeit in Anspruch genommen hatten, dass die Durchführung der optionalen Untertests *Rechnerisches Denken* und *Allgemeines Wissen* entweder aufgrund der Ermüdung nicht zumutbar, oder aufgrund von zeitlichen Verpflichtungen der Testpersonen oder ihrer Eltern nicht möglich waren. Jedenfalls stellen die 28 Personen, die diese Untertests bearbeiteten, eine noch zusätzlich ausgelesene Stichprobe dar, die vor allem die „eher

schnell arbeitenden“ Testpersonen sein dürften, doch ist leider nicht eindeutig eruierbar, welche Gründe letztendlich dazu geführt haben, dass die anderen 13 Personen diese Untertests nicht bearbeiteten, was die Aussagekraft der Ergebnisse stark verringert.

### 5.9.2 Itemtrennschärfe, Itemschwierigkeit

Der Untertest *Rechnerisches Denken* beinhaltet 34 Items, die dichotom verrechnet werden, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt. Tabelle 5.11 zeigt sowohl die Anzahl der Personen, die diese Items bearbeiteten, als auch die relativen Lösungshäufigkeiten und die Trennschärfeindizes über 0,25:

**Tabelle 5.11:** Itemschwierigkeiten und Trennschärfen (> 0,25) des Untertests *Rechnerisches Denken*

Item-Nr. <sup>23</sup>	N	Itemschwierigkeit (rel. Lösungshäufigkeit)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
RD5	1	1		
RD6	1	1		
RD7	1	1		
RD8	1	1		
RD9	10	1		
RD10	10	1		
RD11	11	1		
RD12	28	1		
RD13	28	1		
RD14	28	0,96	0,28	0,29
RD15	28	0,96		
RD16	28	0,96	-0,25	-0,26
RD17	28	1		
RD18	28	1		
RD19	28	0,96		
RD20	28	0,96	0,46	0,32
RD21	28	0,96	0,46	0,32

<sup>23</sup> Die Items 1, 2, 3 und 4 wurden keiner Testperson vorgegeben und scheinen daher hier nicht auf.

RD22	28	0,79	0,40	0,34
RD23	28	0,96	0,46	0,32
RD24	27	0,96	0,35	0,32
RD25	27	0,89		
RD26	27	0,63	0,48	0,49
RD27	27	0,67	0,60	0,62
RD28	27	0,48	0,54	0,56
RD29	26	0,54	0,51	0,50
RD30	22	0,64		
RD31	21	0,24	0,74	0,73
RD32	21	0,33	0,60	0,64
RD33	19	0,21	0,45	0,46
RD34	16	0,19	0,31	0,32

---

Die Items 1 bis 4 wurden gar nicht vorgegeben, die Items 5 bis 21, sowie 23 und 24 wurden entweder von allen Testpersonen gelöst oder nur von einer Testperson nicht gelöst, was bedeutet, dass sie kaum Informationen hinsichtlich der Unterschiede der Leistungsniveaus der Testpersonen liefern.<sup>24</sup>

Die auffällige negative (!) Trennschärfe des Items 16 beruht nur auf einer einzigen Testperson, weshalb dieses Ergebnis unerheblich ist.

Der Bereich der relativ leichten Items (also mit Lösungshäufigkeiten zwischen 0,8 bis 0,5) ist mit fünf Items repräsentiert, von diesen weisen nur vier Items (die Items 22, 26, 27, und 29) Trennschärfen zwischen 0,3 und 0,6 auf, was als zu wenig erscheint, um in

---

<sup>24</sup> Demnach wurden jeder der Testpersonen etwa 12 Items (oder mehr!) vorgegeben, die eine Lösungswahrscheinlichkeit nahe 1 haben. Wenn man außerdem bedenkt, dass bei etwas weniger als der Hälfte der Testpersonen (12 von 28) die Abbruchregel zum Tragen kam, die besagt, dass nach vier aufeinanderfolgenden Items ohne Lösung die Vorgabe abgebrochen wird, so wird klar, dass für die meisten Testpersonen weniger als die Hälfte der vorgegebenen Items in einem angepassten Schwierigkeitsbereich war. Vielmehr war der Großteil zu leicht und – wenigstens bei 12 Testpersonen – waren mindestens vier Items (nämlich die, die zur Abbruchregel geführt haben) zu schwierig, was aber – folgt man der Ausführung zu artifiziell hohen Split-half-Reliabilitäten am Anfang des Kapitels – mitunter zu relativ hohen Korrelationen zwischen den Testteilen und damit zu einigermaßen hohen Reliabilitätskoeffizienten führen kann, ohne dass die Items wirklich die gleiche Fähigkeitsdimension prüfen.

diesem Schwierigkeitsbereich Fähigkeitsdifferenzen der Testpersonen adäquat abbilden zu können.

Ähnliches gilt für den Bereich der schwierigeren Items. So wurden nur fünf Items in dieser Teilstichprobe von weniger als 50 % gelöst, wobei deren Trennschärfen mit Werten zwischen 0,3 und 0,7 durchaus als ausreichend gelten können.

An dieser Stelle muss aber nochmals darauf hingewiesen werden, dass diese Teilstichprobe nur 28 Personen umfasst und die Generalisierbarkeit der Ergebnisse stark eingeschränkt ist.

Eine Berechnung der Reliabilitätskoeffizienten in den drei Altersgruppen erschien aufgrund der kleinen Stichprobengröße nicht sinnvoll (die Altersgruppe 2 hat einen Umfang von nur 7 Personen).

## **5.10 Allgemeines Wissen – Reliabilität, Trennschärfe, Itemschwierigkeit**

### **5.10.1 Vergleich der Reliabilitätsangaben des HAWIK-IV-Manuals mit den empirischen Werten aus der Stichprobe**

Die mittlere Reliabilität laut Manual (über alle Altersgruppen von 6 – 16 Jahren) liegt bei  $r_{tt} = 0,85$ , die Reliabilitäten pro Lebensalter (in Jahren) liegen zwischen 0,76 und 0,90 (Petermann & Petermann, 2007).

Für die Stichprobe der vorliegenden Arbeit wurde eine empirische Split-half-Reliabilität von  $r_{tt} = 0,70$  (sowohl nach Pearson als auch nach Spearman) errechnet. Ergänzend dazu muss aber festgestellt werden, dass nur 24 Testpersonen diesen Untertest bearbeiteten; der Grund dafür ist im vorhergehenden Kapitel erklärt worden.

### **5.10.2 Itemtrennschärfe, Itemschwierigkeit**

Der Untertest *Allgemeines Wissen* beinhaltet 33 Items, die dichotom verrechnet werden, wobei es pro Altersgruppe unterschiedliche Einstiegsitems gibt; In Tabelle 5.12 sind einerseits die Anzahl der Tpn, die diese Items bearbeiteten, als auch die relativen Lösungshäufigkeiten und die Trennschärfeindizes über 0,25 angeführt:

**Tabelle 5.12:** Itemschwierigkeiten und Trennschärfen ( $> 0,25$ ) des Untertests *Allgemeines Wissen*

Item-Nr. <sup>25</sup>	N	Itemschwierigkeit (rel. Lösungshäufigkeit)	Trennschärfe (Pearson)	Trennschärfe (Spearman)
AW5	3	1		
AW6	3	1		
AW7	3	1		
AW8	3	1		
AW9	3	1		
AW10	16	1		
AW11	16	1		
AW12	24	1		
AW13	24	1		
AW14	24	0,75	0,38	0,39
AW15	24	0,92	0,3	0,34
AW16	24	0,87	0,43	0,42
AW17	24	0,96	0,27	0,30
AW18	24	0,83	0,51	0,50
AW19	24	1		
AW20	24	0,42		
AW21	24	0,87	0,51	0,49
AW22	24	0,75	0,35	0,33
AW23	24	0,46	0,58	0,64
AW24	24	0,38	0,5	0,5
AW25	23	0,78	0,46	0,54
AW26	23	0,26	0,59	0,6
AW27	22	0,05		
AW28	19	0,11	0,36	0,39
AW29	19	0,37	0,29	0,30
AW30	18	0,17	0,29	0,30
AW31	12	0,08	0,35	0,35
AW32	9	0,78	0,58	0,57
AW33	9	0		

<sup>25</sup> Die Items 1, 2, 3 und 4 wurden keiner Testperson vorgegeben und scheinen daher hier nicht auf.

Die Items 1 bis 4 wurden gar nicht vorgegeben, die Items 5 bis 13, 16 und 19 wurden von allen Testpersonen gelöst oder nur von einer Testperson nicht gelöst, was bedeutet, dass sie keine oder kaum Informationen hinsichtlich der Unterschiede der Leistungsniveaus der Testpersonen liefern. Weiter vier Items (mit Lösungshäufigkeiten über 0,8) können als ausgesprochen leicht gelten, nur vier Items liegen im Schwierigkeitsbereich zwischen 0,5 und 0,8 (mit Trennschärfen zwischen 0,3 und 0,6), was als recht wenig erscheint, um in diesem Bereich Fähigkeitsdifferenzen der Testpersonen adäquat abbilden zu können. Wieder muss an dieser Stelle aber darauf hingewiesen werden, dass diese Teilstichprobe von 24 Personen ausgelesen ist und die Generalisierbarkeit dieser Kritik stark in Zweifel gezogen werden muss.

Der Bereich der schwierigeren Items ist mit sieben Items mit Trennschärfen zwischen etwa 0,3 und 0,6 vergleichsweise gut repräsentiert.

Eine Berechnung der Reliabilitätskoeffizienten in den drei Altersgruppen erschien aufgrund der kleinen Stichprobengröße nicht sinnvoll.

### **5.11 Reliabilität der Untertests ZST, ZN, BZF und SYS**

Da aufgrund der Art dieser Untertests keine Bestimmung der Reliabilität aus den vorliegenden Daten durchgeführt werden kann, werden die Untertests *Zahlen-Symbol-Test*, *Zahlen Nachsprechen*, *Buchstaben-Zahlen-Folgen* und *Symbolsuche* hier nicht behandelt.

### **5.12 Empirische Split-half-Reliabilität der nachträglich dichotomisierten Untertests GF, AV und WT**

Bezüglich der dreikategoriellen Verrechnung der drei Untertests *Gemeinsamkeiten Finden*, *Allgemeines Verständnis* und *Wortschatz-Test* wurden bereits mehrere Kritikpunkte erläutert, die die inhaltliche Genauigkeit und Interpretierbarkeit im Sinne der Validität, die Skalierung und die praktische Handhabung betreffen. Aus diesem Grund wird an dieser Stelle der Frage nachgegangen, ob der *Gewinn* einer solchen dreikategoriellen Verrechnung - und das ist m.E. nur ein Zuwachs an Messgenauigkeit – überhaupt in relevantem Ausmaß gegeben ist.

Um der Frage nachzugehen, ob auch eine dichotome Verrechnung zu ähnlicher Messgenauigkeit führen könnte wie die dreikategorielle Verrechnung, wurden die Ergebnisse der drei Untertests *Gemeinsamkeiten Finden*, *Wortschatz-Test* und



*Allgemeines Verständnis* nachträglich dichotomisiert. Dabei wurden in der Version „dichotom\_leicht“ sowohl die Ein-Punkt-Antworten als auch die Zwei-Punkt-Antworten gleichermaßen als gelöst gewertet, wohingegen in der Version „dichotom\_schwierig“ nur die Zwei-Punkt-Antworten als Lösung verrechnet wurden, die Ein-Punkt-Antworten (im Sinne der Aussage „halb gelöst gibt es nicht“) allerdings als „nicht-gelöst“.

Da schon in der dreikategoriellen Verrechnungsweise z.T. zu wenige Items vorhanden sind, die schwierig genug sind, um im oberen Leistungsbereich gut zu differenzieren, ist davon auszugehen, dass dieses Problem in der Version „leicht“ noch deutlicher zum Tragen kommt. Dennoch wurden für alle Versionen Split-half-Reliabilitäten berechnet und in folgender Tabelle (5.13) denen der ursprünglichen dreikategoriellen Versionen gegenübergestellt.

**Tabelle 5.13:** Split-half-Reliabilitätskoeffizienten der ursprünglichen und der nachträglich dichotomisierten Skalen

		dreikategoriell	dichotom_ leicht	dichotom_ schwierig
Gemeinsamkeiten Finden	Pearson	0,78	0,76	0,61
	Spearman	0,76	0,75	0,62
Wortschatz-Test	Pearson	0,81	0,81	0,75
	Spearman	0,74	0,70	0,73
Allgemeines Verständnis	Pearson	0,89	0,86	0,83
	Spearman	0,88	0,87	0,82

Deskriptiv zeigt sich, dass die Reliabilitätskoeffizienten der Dichotomisierung „leicht“ nicht oder kaum geringer sind als die der dreikategoriellen Versionen. Durch die Dichotomisierung „schwierig“ kommt es aber deskriptiv zu einer geringeren Split-half-Reliabilität beim Untertest *Gemeinsamkeiten Finden*, beim Untertest *Wortschatz-Test* dagegen führt auch sie zu keiner Verminderung des Reliabilitätskoeffizienten und auch beim Untertest *Allgemeines Verständnis* zeigt sich nur eine marginale Verminderung.

Um der Frage nachzugehen, inwiefern in den Untertest-Versionen „dichotom\_leicht“ genügend schwierige Items zur Verfügung stehen, werden folgend die Itemkennwerte (Schwierigkeiten und Trennschärfen) dargestellt.

### 5.12.1 Itemkennwerte des dichotomisierten Untertests *Gemeinsamkeiten Finden*

**Tabelle 5.14:** Itemschwierigkeiten der dreikategoriellen und dichotomisierten Skalen des Untertests *Gemeinsamkeiten Finden*

Item-Nr.	N	Itemschwierigkeit		
		dreikategoriell	dichotom_leicht	dichotom_schwierig
GF1	5	1	1	1
GF2	5	1	1	1
GF3	26	0,96	1	0,93
GF4	26	1	1	1
GF5	41	0,95	0,98	0,93
GF6	41	0,94	1	0,88
GF7	41	1	1	1
GF8	41	0,84	0,85	0,83
GF9	41	0,84	0,9	0,78
GF10	41	0,77	0,85	0,68
GF11	41	0,77	0,85	0,68
GF12	41	0,92	0,98	0,85
GF13	41	0,88	0,95	0,80
GF14	41	0,60	0,66	0,54
GF15	41	0,71	0,80	0,61
GF16	41	0,57	0,71	0,44
GF17	41	0,56	0,61	0,51
GF18	41	0,48	0,54	0,41
GF19	40	0,38	0,63	0,13
GF20	40	0,41	0,55	0,28
GF21	39	0,42	0,56	0,28
GF22	38	0,28	0,47	0,08
GF23	38	0,21	0,32	0,11

War für die ursprüngliche, dreikategorielle Version des Untertests *Gemeinsamkeiten Finden* noch zu bemängeln, dass es nur zwei Items mit einer Itemschwierigkeit unter 0,3 und nur vier weitere Items mit einer Itemschwierigkeit unter 0,5 gab, so ist dies für die Version *GF\_dichotom\_leicht* noch deutlich problematischer (s. Tabelle 5.14): hier zeigen sich gar keine Items mit einer Schwierigkeit unter 0,3 und nur zwei Items mit einer Schwierigkeit (relativen Lösungshäufigkeit) unter 0,5 (und zumindest sechs weitere mit einer Itemschwierigkeit unter 0,7, wodurch wenigstens im mittleren Leistungsbereich eine Differenzierung möglich wäre). Die Dichotomisierung „leicht“ wäre für diesen Untertest also nicht sinnvoll.

In der Version *GF\_dichotom\_schwierig* erreichten in der vorliegenden Stichprobe zumindest fünf Items eine Itemschwierigkeit unter 0,3 und zwei weitere eine Itemschwierigkeit unter 0,5, was den oberen Leistungsbereich besser abbilden könnte. Dennoch muss angemerkt werden, dass die Split-half-Reliabilität dieser Version mit 0,61 (nach Pearson) deutlich zu gering ausfällt. Für diesen Untertest scheint eine einfache – das heißt, an den vorliegenden Items orientierte – dichotome Version nicht sinnvoll.

### 5.12.2 Itemkennwerte des dichotomisierten Untertests *Wortschatz-Test*

Tabelle 5.15 gibt die Itemschwierigkeiten der dreikategoriellen und dichotomisierten Skalen des Wortschatz-Tests wieder.

**Tabelle 5.15:** Itemschwierigkeiten der dreikategoriellen und dichotomisierten Skalen des *Wortschatz-Tests*

Item-Nr.	N	Itemschwierigkeit		
		dreikategoriell	dichotom_leicht	dichotom_schwierig
WT5	5	0,90	1	0,8
WT6	5	1	1	1
WT7	27	1	1	1
WT8	27	1	1	1
WT9	41	0,99	1	0,98
WT10	41	0,99	1	0,98
WT11	41	1	1	1

WT12	41	0,98	0,98	0,98
WT13	41	0,93	0,98	0,88
WT14	41	0,93	1	0,85
WT15	41	0,93	1	0,85
WT16	41	0,94	0,98	0,9
WT17	41	0,98	1	0,95
WT18	41	0,95	0,95	0,95
WT19	41	0,83	0,83	0,83
WT20	41	0,92	0,93	0,90
WT21	41	0,78	0,88	0,68
WT22	41	0,84	0,88	0,80
WT23	41	0,61	0,76	0,46
WT24	41	0,55	0,59	0,51
WT25	41	0,65	0,73	0,56
WT26	41	0,70	0,71	0,68
WT27	41	0,84	0,95	0,73
WT28	40	0,66	0,75	0,58
WT29	40	0,69	0,85	0,53
WT30	40	0,56	0,63	0,5
WT31	40	0,85	0,88	0,83
WT32	39	0,41	0,64	0,18
WT33	39	0,71	0,82	0,59
WT34	39	0,46	0,49	0,44
WT35	39	0,17	0,18	0,15
WT36	39	0,60	0,77	0,44

---

Der *Wortschatz-Test* weist in der Überprüfung anhand der vorliegenden Stichprobendaten in der ursprünglichen, dreikategoriellen Version nur *ein* Item mit einer Schwierigkeit unter 0,3 und nur zwei weitere Items mit einer Itemschwierigkeit unter 0,5 auf, weshalb von starken Deckeneffekten auszugehen ist. In der Version

*WT\_dichotom\_leicht* ist das Ergebnis noch problematischer (ein Item mit der Schwierigkeit unter 0,3, nur ein weiteres unter 0,5).

In der Version *WT\_dichotom\_schwierig* dagegen zeigt sich ein etwas positiveres Bild: Hier weisen zwei Items eine Schwierigkeit unter 0,3 auf und drei weitere eine Itemschwierigkeit unter 0,5 (und ein weiteres mit exakt 0,5), was zwar immer noch zu wenig erscheint, um ausreichend gut im oberen Leistungsbereich differenzieren zu können, aber zumindest besser, als in der dreikategoriellen Version.

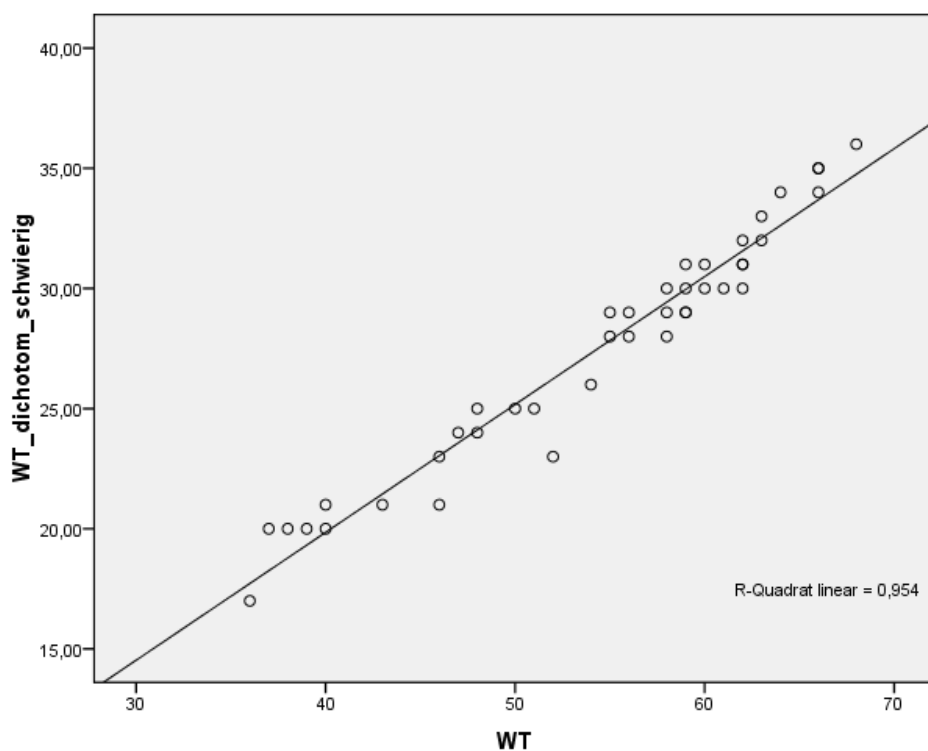
Die folgende Tabelle 5.16 listet die Trennschärfen derjenigen Items auf, die wenigsten in einer der beiden Versionen (dreikategoriell bzw. *dichotom\_schwierig*) eine Lösungshäufigkeit unter 0,6 aufweisen. Es zeigen sich leicht höhere Lösungshäufigkeiten in der dreikategoriellen Version, also eine etwas größere Anzahl schwierigerer Items in der dichotomisierten (*schwierig*) Version bei gleichzeitig etwas niedrigeren Trennschärfen (eine vollständige Tabelle A.1 aller Items ist im Anhang zu finden).

**Tabelle 5.16:** Trennschärfen der Items mit Itemschwierigkeiten unter 0,6 der dreikategoriellen und der dichotomisierten („*schwierig*“) Skala des Untertests *Wortschatz-Test*

Item-Nr.	N	Itemschwierigkeit		Trennschärfe nach Pearson		Trennschärfe nach Spearman	
		drei-kategoriell	dichotom_schwierig	drei-kategoriell	dichotom_schwierig	drei-kategoriell	dichotom_schwierig
WT23	41	0,61	0,46	0,46	0,36	0,45	0,35
WT24	41	0,55	0,51	0,75	0,78	0,76	0,79
WT25	41	0,65	0,56	0,62	0,45	0,56	0,42
WT28	40	0,66	0,58	0,64	0,49	0,57	0,45
WT29	40	0,69	0,53	0,59	0,36	0,48	0,33
WT30	40	0,56	0,50	0,64	0,61	0,67	0,64
WT32	39	0,41	0,18	0,45	0,49	0,51	0,51
WT33	39	0,71	0,59	0,57	0,42	0,43	0,36
WT34	39	0,46	0,44	0,37	0,33	0,34	0,32
WT35	39	0,17	0,15	0,41	0,43	0,47	0,46
WT36	39	0,60	0,44	0,58	0,49	0,65	0,51

Die Frage, ob die Ergebnisse der Testpersonen in der Skala *WT\_dichotom\_schwierig* denjenigen entsprechen, die durch die dreikategorielle Version erzielt wurden, lässt sich teilweise durch den statistischen Zusammenhang zwischen diesen Ergebnissen darstellen. Eine Berechnung der Korrelation zwischen den dreikategoriellen Scores und den dichotomisierten (schwierig) Scores aller 41 Testpersonen ergab mit einem Wert von  $r = \rho = 0,98$  ein Ergebnis, das darauf schließen lässt, dass ein in dieser Form dichotomisierter Untertest gleich messgenau zu fast den gleichen Ergebnissen (hinsichtlich der Positionierung der Testpersonen zueinander) käme wie der dreikategorielle Untertest, ohne die beschriebenen Probleme der Skalierung und der praktischen Testvorgabe aufzuweisen (siehe Abbildung 5.4).

**Abbildung 5.4:** Streudiagramm der *Wortschatz-Test*-Scores der Testpersonen in der dreikategoriellen Version und in der Version „dichotom-schwierig“



### 5.12.3 Itemkennwerte des dichotomisierten Untertests *Allgemeines Verständnis*

Folgende Tabelle gibt die Itemschwierigkeiten der dreikategoriellen und dichotomisierten Skalen des Untertests *Allgemeines Verständnis* wieder.

**Tabelle 5.17:** Itemschwierigkeiten der dreikategoriellen und dichotomisierten Skalen des Untertests *Allgemeines Verständnis*

Item-Nr.	N	Itemschwierigkeit (durchschnittliche Punkte / 2)		
		dreikategoriell	dichotom_leicht	dichotom_schwierig
AV1	4	1	1	1
AV2	4	1	1	1
AV3	26	0,98	1	0,96
AV4	26	0,92	0,96	0,88
AV5	41	0,99	1	0,98
AV6	41	1	1	1
AV7	41	0,91	0,93	0,9
AV8	41	0,82	0,88	0,76
AV9	41	0,79	1	0,59
AV10	41	0,80	0,98	0,63
AV11	41	0,61	0,8	0,41
AV12	41	0,84	0,88	0,80
AV13	41	0,35	0,44	0,27
AV14	41	0,50	0,76	0,24
AV15	41	0,66	0,66	0,66
AV16	39	0,53	0,74	0,31
AV17	38	0,67	0,71	0,63
AV18	38	0,28	0,47	0,08
AV19	37	0,31	0,51	0,11
AV20	34	0,32	0,53	0,12
AV21	30	0,27	0,37	0,17
AV21	30	0,27	0,37	0,17

War für die ursprüngliche, dreikategorielle Version des Untertests *Allgemeines Verständnis* noch zu bemängeln, dass es nur zwei Items mit einer Itemschwierigkeit unter 0,3 und nur drei weitere Items mit einer Itemschwierigkeit unter 0,5 gibt, so weist die Version *AV\_dichotom\_leicht* gar keine Items mit einer Schwierigkeit unter 0,3 und nur drei mit einer Itemschwierigkeit unter 0,5 auf, weswegen diese Dichotomisierung (leicht) zu einer sehr mangelhaften Differenzierung im oberen Leistungsbereich führen würde.

Die Version *AV\_dichotom\_schwierig* dagegen weist in der vorliegenden Stichprobe sechs Items mit einer Schwierigkeit unter 0,3 und zwei weitere mit einer Itemschwierigkeit unter 0,5 auf, womit diese Version den oberen Leistungsbereich besser differenzieren könnte als die dreikategorielle.

Tabelle 5.18 listet die Trennschärfen derjenigen Items auf, die wenigsten in einer der beiden Versionen (dreikategoriell bzw. *dichotom\_schwierig*) eine Lösungshäufigkeit unter 0,6 aufweisen. Es zeigt sich eine größere Anzahl schwierigerer Items in der dichotomisierten (*schwierig*) Version bei gleichzeitig etwas niedrigeren Trennschärfen (eine vollständige Tabelle A.2 aller Items ist im Anhang zu finden).

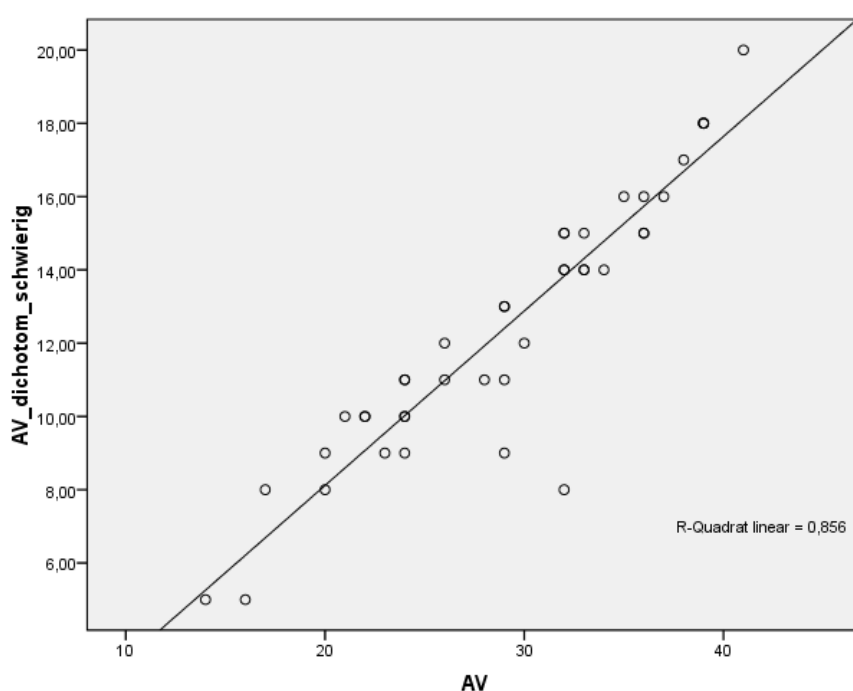
**Tabelle 5.18:** Trennschärfen der Items mit Itemschwierigkeiten unter 0,6 der dreikategoriellen und der dichotomisierten („*schwierig*“) Skala des Untertests *Allgemeines Verständnis*

Item-Nr.	N	Itemschwierigkeit		Trennschärfe nach Pearson		Trennschärfe nach Spearman	
		drei-kategoriell	dichotom_schwierig	drei-kategoriell	dichotom_schwierig	drei-kategoriell	dichotom_schwierig
AV9	41	0,79	0,59	0,33	0,36	0,27	0,36
AV11	41	0,61	0,41	0,46	0,49	0,54	0,52
AV13	41	0,35	0,27	0,58	0,66	0,63	0,67
AV14	41	0,50	0,24	0,61	0,64	0,73	0,65
AV16	39	0,53	0,31	0,42	0,16	0,40	0,17
AV18	38	0,28	0,08	0,57	0,45	0,64	0,44
AV19	37	0,31	0,11	0,59	0,30	0,69	0,27
AV20	34	0,32	0,12	0,65	0,32	0,65	0,29
AV21	30	0,27	0,17	0,63	0,42	0,71	0,48



Eine Berechnung der Korrelation zwischen den dreikategoriellen Scores und den dichotomisierten (schwierig) Scores aller 41 Testpersonen ergab mit einem Wert von  $r = 0,93$  bzw.  $\rho = 0,92$  (siehe auch Abb. 5.5) auch ein Ergebnis, das darauf schließen lässt, dass ein in dieser Form dichotomisierter Untertest gleich messgenau zu fast den gleichen Ergebnissen (hinsichtlich der Positionierung der Testpersonen zueinander) käme wie der dreikategorielle Untertest, ohne die beschriebenen Probleme der Skalierung und der praktischen Testvorgabe aufzuweisen.

**Abbildung 5.5:** Streudiagramm der Scores der Testpersonen in der dreikategoriellen Version und in der Version „dichotom-schwierig“ des Untertests *Allgemeines Verständnis*



### 5.13 Diskussion und Zusammenfassung der Ergebnisse zur Reliabilität

In diesem Kapitel wurde zu Beginn gezeigt, dass nach der Split-half-Methode berechnete Reliabilitätskoeffizienten sehr empfindlich gegenüber Verzerrungen sind, die entstehen, wenn viele Items eine deutlich zu niedrige oder deutlich zu hohe Lösungshäufigkeit in der Stichprobe haben, mit der die Reliabilität empirisch ermittelt werden soll. Alleine schon aus diesem Grund wurden auch die Lösungshäufigkeiten und die Trennschärfe auf Itemebene dargestellt; darüber hinaus war das notwendig, um abschätzen zu können, inwiefern die jeweiligen Untertests in der Lage sind, in den verschiedenen Leistungsbereichen gut genug zu differenzieren.

Es zeigten sich in mehreren Untertests eine deutlich zu geringe Reliabilität bzw. eine nur mangelhafte Differenzierungsfähigkeit im oberen Leistungsbereich, was auf sogenannte Deckeneffekte hinausläuft. Dies betrifft zumindest die Untertests *Mosaik-Test* (die Power-Komponente, nicht aber die Speedkomponente), *Bildkonzepte* und *Wortschatz-Test*. Da diese Untertests sich aufgrund der mangelnden Messgenauigkeit (bzw. Deckeneffekte) daher nicht für eine Profilinterpretation eignen, ist dieses Ergebnis auch als weitere Einschränkung der Verwendbarkeit des HAWIK-IV für die Hochbegabungsdiagnostik nach dem *Wiener Diagnosemodells* zum *Hochleistungspotenzial* zu werten.

Aufgrund der im Kapitel 4 diskutierten Kritikpunkte an der dreikategoriellen Verrechnung wurde untersucht, inwiefern nachträglich dichotomisierte Versionen der drei Untertests *Gemeinsamkeiten Finden*, *Wortschatz-Test* und *Allgemeines Verständnis* zu vergleichbar messgenauen Ergebnissen führen würden. Die Ergebnisse legen nahe, dass dies für den Untertest *Gemeinsamkeiten Finden* nicht gelingen würde. In der Version „dichotom\_schwierig“ aber würde dies für den *Wortschatz-Test* sogar zusätzliche Vorteile bringen, weil dadurch möglicherweise auch der Deckeneffekt zu verringern wäre, und auch für den Untertest *Allgemeines Verständnis* scheint eine Dichotomisierung sinnvoll möglich zu sein. Ob dies auch in anderen Stichproben gilt, ist aus diesen Ergebnissen aber nicht abzuleiten.

## 6 FAIRNESS

„Ein Test erfüllt das Gütekriterium *Fairness*, wenn die resultierenden Testwerte zu keiner systematischen Diskriminierung bestimmter Testpersonen zum Beispiel aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit führen.“ (Kubinger, 2006, S. 118)

Definitionen von *Fairness* als Gütekriterium eines psychologisch-diagnostischen Verfahrens beziehen sich regelmäßig auf die Diskriminierung bestimmter Testpersonen aufgrund ihrer Gruppenzugehörigkeit.

Im alltäglichen Sprachgebrauch bezieht sich *Fairness* schlicht darauf, dass Individuen nicht benachteiligt werden: üblicherweise werden sowohl Verstöße gegen bestehende Regeln (im Sport, aber auch in der Benotung in der Schule etc.) als unfair bezeichnet, als auch Verstöße gegen den „Geist eines Gesetzes“: ein Beispiel dafür wäre, wenn ein Lehrer einen Schüler zur Prüfung ausschließlich über die Themen befragt, die durchgenommen wurden, als dieser Schüler krank war. Solche Vorgehensweisen werden von der Mehrzahl unmittelbar als *unfair* oder *ungerecht* empfunden. So findet sich im Internet-Wörterbuch „Free Dictionary“ auch folgende Definition: „Fairness: [...] gerechtes und anständiges Verhalten“ (<http://de.thefreedictionary.com/Fairness>, 20.10.08). Für das Verständnis von „unfair“ ist es unerheblich, ob dabei ethnische oder andere Gruppenzugehörigkeiten eine Rolle spielen, oder ob einzelne Personen nur aufgrund ihrer individuellen Eigenart benachteiligt werden.

*Fairness* kann aber auch als „Angemessenheit“ verstanden werden (<http://de.wikipedia.org/wiki/Fairness>, 20.10.08); auch in diesem Sinne wird es im Rahmen psychologischen Diagnostizierens verwendet: „Verrechnungsfairness“ ist gegeben, wenn die Verrechnungsvorschriften zu Testwerten führen, die die empirischen Verhaltensrelationen adäquat wiedergeben; „Verrechnungsfairness“ ist somit ein wesentlicher Aspekt des Gütekriteriums *Skalierung* (frei nach Kubinger, 2006).

Die folgenden Kritikpunkte an den Verrechnungs- und Vorgabevorschriften des HAWIK-IV orientieren sich weniger an der Definition von *Fairness* als Gütekriterium, sondern mehr am alltäglichen Verständnis von *Fairness*, und zwar deswegen, weil nicht die systematische Benachteiligung von bestimmten Gruppen aufgezeigt werden soll,

sondern die Benachteiligung von Personen mit bestimmten *Persönlichkeitseigenschaften*.

## 6.1 Die Abbruchregeln

Mit Ausnahme der beiden zum Index *Verarbeitungsgeschwindigkeit* gehörenden Untertests *Symbolsuche* und *Zahlensymbol-Test* beinhalten die Vorgaben zu allen anderen hier vorgestellten Untertests ein Abbruchkriterium: dieses ist erreicht, wenn je nach Untertest drei bis fünf Items hintereinander nicht gelöst wurden (bzw. vier von fünf Items nicht gelöst wurden). Diese Vorgangsweise stützt sich offensichtlich auf die Annahme, die Items wären nach ihrer Schwierigkeit aufsteigend gereiht, denn nur dann kann angenommen werden, dass eine Testperson, welche die bisherigen Items nicht lösen konnte, die noch folgenden (also schwierigeren) Aufgaben auch nicht lösen kann.

Abgesehen davon, dass die Annahme, die Items wären (streng monoton) nach aufsteigender Schwierigkeit gereiht, von den Autoren nicht belegt (und möglicherweise nicht geprüft) wurde, würde diese Vorgangsweise auch bei einer stetigen Zunahme der Itemschwierigkeit zu Benachteiligungen einzelner Testpersonen führen. Denn weshalb sollte eine Testperson, welche beispielsweise die Items 10, 11, 12, 13, 15 und 16 nicht lösen konnte, eher in der Lage sein, das Item 17 zu lösen, als eine Testperson, welche die Items 12, 13, 14, 15 und 16 nicht lösen konnte? Bei erster Person wäre beispielsweise das Abbruchkriterium „fünf hintereinander nicht gelöste Items“ nicht erfüllt; sie würde daher noch weitere Items vorgelegt bekommen. Die zweite Person dagegen würde diese Chance nicht bekommen (eben weil sie fünf *aufeinanderfolgende* Items nicht lösen konnte), und zwar, obwohl sie bis dahin sogar *mehr* Items gelöst hätte, als die erste!

Dies ist insofern bemerkenswert, als nur im Falle des Abbruchkriteriums die Leistung der letzten drei, vier oder fünf Items als adäquate Statistik der Fähigkeit der Testperson herangezogen wird, während für die eigentlichen Messung dieser Fähigkeit (also die Umrechnung in normierte Werte) die *insgesamt erworbenen* Punkte (bzw. die Anzahl gelöster Items) als Maßzahl gilt.

Gegen diese letztere Verrechnung könnte ja (im Sinne der Skalierung) durchaus argumentiert werden: So erscheint eine Berechnung der Fähigkeitsparameter (z.B. Wertpunkte) anhand des Scores zwar üblich zu sein, dennoch kann nicht verständlich argumentiert werden, weshalb das Nicht-lösen relativ einfacher Items in gleicher Weise verrechnet wird, wie das Nicht-Lösen eines schwierigen Items (vgl. Kubinger, 2006). Ob

die Anzahl der gelösten Items ohne Einbezug der Information, die sich daraus ergibt, *welche* Items denn nicht gelöst werden konnten und welche schon, wirklich eine erschöpfende Statistik der Leistung einer Testperson ist oder nicht, kann nämlich ohne eine entsprechende testtheoretische Prüfung ohnehin nicht klar ausgesagt werden<sup>26</sup> (Kubinger, 2006).

Unabhängig von diesen Überlegungen lässt sich konstatieren: diese Vorgangsweise (nämlich die Heranziehung der *Summe* der je Untertest erworbenen Punkte als Maß für die jeweilige Fähigkeit) ist im HAWIK-IV jedenfalls der *gültige* Modus für die Umrechnung in Normwerte. Wenn aber die Testautoren davon ausgehen, dass die Anzahl der gelösten Aufgaben ein korrektes Maß der Testleistung ist, ist unverständlich, weshalb im Falle des Abbruchkriteriums ein anderes Maß dafür herangezogen wird. In den folgenden Abschnitten wird gezeigt werden, dass in vielen Fällen ein Widerspruch zwischen diesen beiden Vorgangsweisen besteht, und zwar zwischen der Entscheidung zu einem etwaigen Abbruch, die aufgrund der letzten drei bis fünf Items getroffen wird (wie es die Abbruchkriterien vorgeben), und der Einschätzung der Testleistung, wie sie sich aus der Summe der bisher erworbenen Punkte (bzw. der Anzahl gelöster Items) ergibt. Es lässt sich nämlich zeigen, dass es bei Testpersonen zu einem Abbruch kommt, die bis zum Abbruch *gleich viele* oder sogar *mehr* Items lösen konnten wie andere Testpersonen, bei denen es *nicht* zum Abbruch kommt, und zwar nur wegen der unterschiedlichen *Reihenfolge* der gelösten bzw. nicht-gelösten Items.

Davor gilt es aber noch zu argumentieren, weshalb diese Untersuchung unter dem Punkt „Fairness“ abgehandelt wird, scheint es doch vor allem eine Einschränkung der *Messgenauigkeit* zu bedeuten, wenn durch die *zufällige* Reihenfolge der gelösten und nicht gelösten Items mitbestimmt wird, ob noch weitere Items vorgelegt werden oder nicht<sup>27</sup>. Es erscheint aber fraglich, ob diese Reihenfolgen nicht – abgesehen vom Zufall – auch durch Persönlichkeitseigenschaften systematisch beeinflusst werden. Bedenkt man nämlich, dass Testpersonen in den meisten Fällen wissen oder zumindest ahnen, ob

---

<sup>26</sup> Diese Aussage kann dahingehend präzisiert werden, dass das *dichotome logistische Modell* von Rasch (oder eine monotone Transformation davon) notwendigerweise gelten muss, damit die Anzahl der gelösten Items ein verrechnungsfaires Maß der Testleistung darstellt (Kubinger, 2006). Dies wurde aber beim HAWIK-IV nicht geprüft.

<sup>27</sup> Es ist offensichtlich auch ein Problem der *Skalierung*, weil die Verrechnungsvorschrift bei den Abbruchkriterien („die letzten drei/vier/fünf Items nicht gelöst“) Zweifel daran nähren, ob die Verrechnung der Gesamtpunkteanzahl (unabhängig davon, *welche* Items gelöst wurden) als Maß der erbrachten Leistung angemessen ist (siehe vorhergehende Fußnote).

sie die letzten Aufgaben lösen konnten oder nicht, so wird klar, dass es bei diesen nach einigen nicht gelösten Items zu Gefühlen der Unsicherheit, Frustration und zu einer niedrigeren Erwartung, die nächsten Items lösen zu können, kommen kann. Diese Gefühle und Gedanken können Auswirkungen auf die Anstrengungsbereitschaft und Leistungsmotivation gegenüber den nächsten Items haben. Dies kann die Bearbeitungsdauer oder die Konzentration verringern, da durch die Misserfolge Selbstzweifel oder Ängste ausgelöst werden können. Diese negativen Auswirkungen der Misserfolgserlebnisse sind stark abhängig von bestimmten Persönlichkeitseigenschaften, wie Selbstwirksamkeitsüberzeugungen, Attributionsstil, Willenskontrolle, Ehrgeiz und Anstrengungsbereitschaft.

Aufgrund dieser Überlegungen schien es sinnvoll, nach systematischen Benachteiligungen von Personen mit bestimmten Persönlichkeitseigenschaften zu suchen, wobei dafür einige der (ohnehin) im Rahmen der Testungen erhobenen Persönlichkeitsvariablen zur Anwendung kamen.

## **6.2 Zusammenhänge der durchgeführten Abbrüche mit Persönlichkeitseigenschaften**

Für die Untersuchung, inwieweit die Frühzeitigkeit der Abbrüche mit Persönlichkeitseigenschaften zusammenhängen, wurden folgende, inhaltlich in Frage kommende Variablen verwendet:

Aus den Variablen des PFK 9-14: *Fehlende Willenskontrolle* (VS2), *Schulischer Ehrgeiz* (MO3), *Selbsterleben von Impulsivität* (SB3), sowie der Faktor zweiter Ordnung *Aktives Engagement versus selbstzweiflerischer Rückzug* (F\_IIO\_3). Außerdem die Testleiter-Ratings zur Ausdauer, Anstrengungsbereitschaft und Frustrationstoleranz (sowohl im schulischen Kontext als auch in der Testsituation) sowie zur Selbstüberzeugung und zum Reaktionstypus (Helplessness vs. Mastery).

Für jede Testperson wurde als Maß für die Frühzeitigkeit der Abbrüche die *Gesamtanzahl der(nach dem Abbruch) nicht bearbeiteten Items*<sup>28</sup> berechnet und mit den zuvor erwähnten Persönlichkeitsfaktoren korreliert. Da die Anzahl der nicht-bearbeiteten

---

<sup>28</sup> in allen Untertests, aber ohne *Allgemeines Wissen* und *Rechnerisches Denken*, die ja nur von einem Teil der Stichprobe bearbeitet wurden

Items naturgemäß mit dem Alter korreliert ist, wurde das Alter als Kontrollvariable einberechnet.

Tabelle 6.1 gibt die Korrelationskoeffizienten der absoluten Höhe nach geordnet wieder:

**Tabelle 6.1:** partielle Korrelationen (Kontrollvariable: Alter) der Anzahl der (nach dem Abbruch) nicht mehr bearbeiteten Items mit ausgewählten Persönlichkeitsvariablen

	Korrelations- koeffizient	Signifikanz <sup>29</sup> (zweiseitig)	N
Anstrengungsbereitschaft in der Schule im Sinne von Ehrgeiz und Lernmotivation (Rating)	-0,38	0,016	41
MO3 Schulischer Ehrgeiz (PFK 9-14)	-0,38	0,034	33
Ausdauer/Penetranz bei Aufgabenbearbeitung im schulischen Kontext (Rating)	-0,34	0,031	41
SB3 Selbsterleben von Impulsivität (PFK 9-14)	0,34	0,062	32
Selbstüberzeugung (Rating)	-0,32	0,047	41
VS2 Fehlende Willenskontrolle (PFK 9-14)	0,20	0,286	32
F_IIO_3 Aktives Engagement versus selbstzweiflerischer Rückzug (PFK 9-14)	-0,20	0,41	17
Reaktionstypus nach Dweck: Helplessness vs. Mastery (Mastery=2, Helplessness=0, unbestimmt=1)	-0,19	0,245	41
Frustrationstoleranz als Reaktion auf Misserfolgserlebnisse in der Testsituation	-0,16	0,331	41
SB1 Selbstüberzeugung (PFK 9-14)	-0,15	0,417	32
Frustrationstoleranz als Reaktion auf Misserfolgserlebnisse in der Schule und sonst	-0,10	0,525	41

Die höchsten Zusammenhänge bestehen zu *Anstrengungsbereitschaft in der Schule im Sinne von Ehrgeiz* und zur Skala *MO3 Schulischer Ehrgeiz* mit einem partiellen Korrelationskoeffizienten von jeweils -0,38. Beide Zusammenhänge weisen in die

<sup>29</sup> Da es sich um eine rein explorative Auswertung handelt, die erst im Nachhinein an den bereits erhobenen Daten angewandt wurde, wird hier auf die Festsetzung eines expliziten Signifikanzniveaus und der vorherigen Festsetzung einer als bedeutsam angesehenen Effektgröße verzichtet, um nicht den Eindruck einer statistischen Hypothesenprüfung zu erwecken.

vermutete Richtung: Ein hohes Ausmaß an Anstrengungsbereitschaft und Ehrgeiz geht mit einer geringen Anzahl nichtbearbeiteter Items, also relativ späten Testabbrüchen einher.

Um die Ergebnisse der Korrelation besser abzusichern und vor allem die Größe des Effekts besser abschätzen zu können, wurden für die drei (von den fallführenden Psychologen auf Ratingskalen eingeschätzten) Persönlichkeitsvariablen *Anstrengungsbereitschaft in der Schule im Sinne von Ehrgeiz und Lernmotivation*, *Ausdauer bei Aufgabenbearbeitung im schulischen Kontext* und *Selbstüberzeugung* drei univariate Varianzanalysen berechnet, wobei das Alter als Kovariate miteinbezogen wurde. Abhängige Variable war wiederum die Gesamtzahl der nichtbearbeiteten Items. Folgend die Ergebnisse:

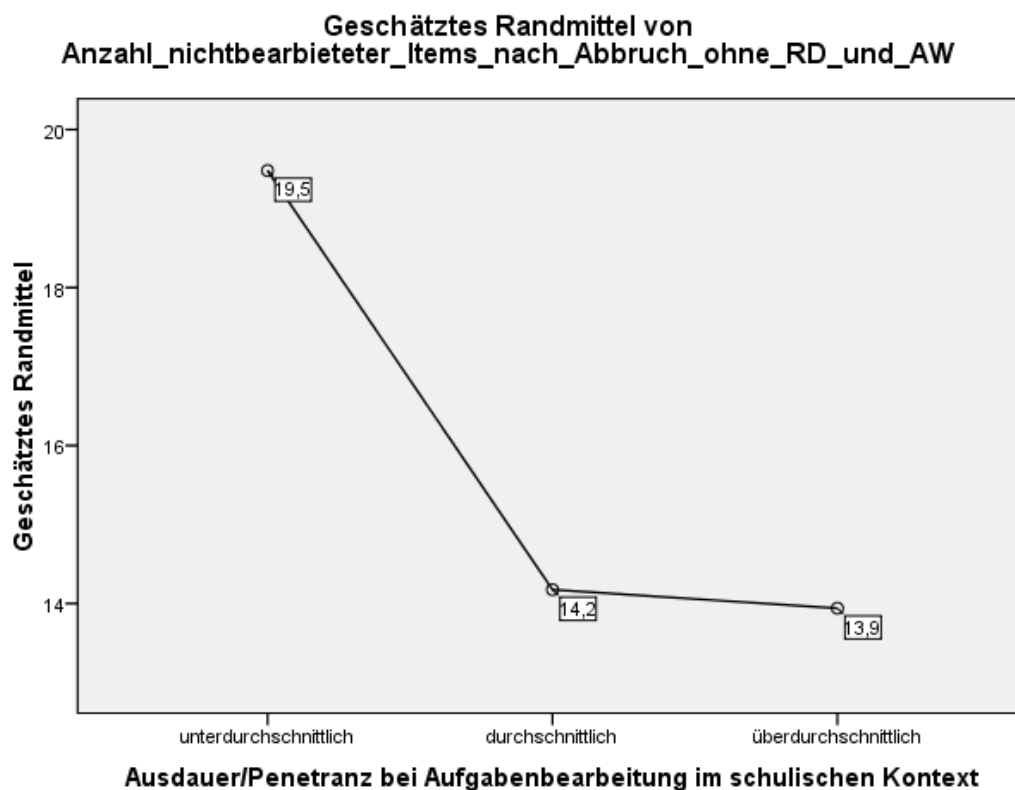
#### 6.2.1 Univariate Varianzanalyse - Ausdauer im schulischen Kontext

Die Anzahl der Personen in den Gruppen der unterschiedlichen Ausprägung von *Ausdauer im schulischen Kontext* beträgt in der Gruppe „unterdurchschnittlich“ 12, durchschnittlich 17 und in der Gruppe „überdurchschnittlich“ 12; Gesamt:  $n = 41$ .

Es zeigt sich zusätzlich zum Effekt des Alters auf die Anzahl der (wegen des Abbruchkriteriums) nicht bearbeiteten Items ein Effekt von *Ausdauer im schulischen Kontext* von  $\eta^2 = 0,156$  ( $p = 0,043$ ). Die (unter Einbezug der Kovariate) geschätzten Randmittel der Anzahl nicht bearbeiteter Items liegen bei den *unterdurchschnittlich ausdauernden* bei 19,5 bei den *durchschnittlich ausdauernden* bei 14,2 und bei den *überdurchschnittlich ausdauernden* bei 13,9 (s. Abb. 6.1). Die paarweisen Vergleiche zeigen eine Differenz von etwa 5,4 *nicht*-bearbeiteten Items zwischen den *unterdurchschnittlich* ausdauernden und den anderen beiden Gruppen ( $p = 0,077$  bzw.  $p = 0,091$  nach Anpassung für Mehrfachvergleiche nach Bonferroni).



**Abbildung 6.1:** Geschätzte Randmittel der univariaten Varianzanalyse der Anzahl nichtbearbeiteter Items, Faktor: *Ausdauer im schulischen Kontext*, Kovariate: Alter



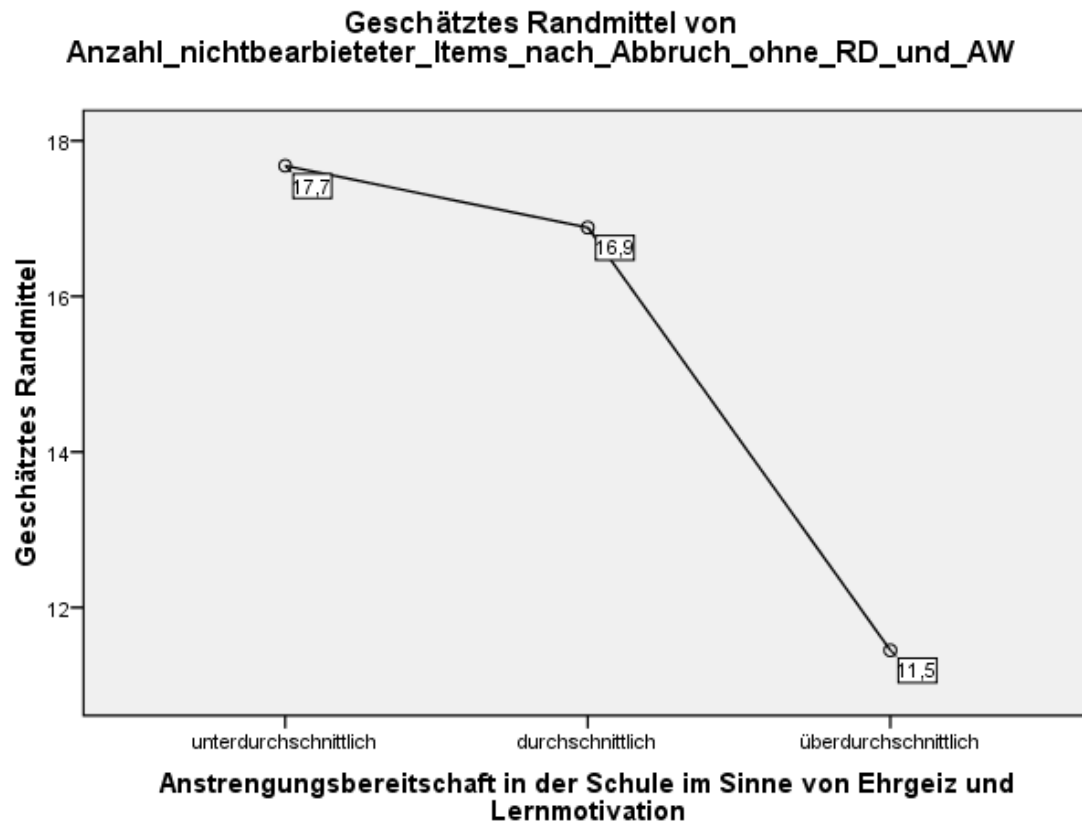
### 6.2.2 Univariate Varianzanalyse - Anstrengungsbereitschaft in der Schule

Die Anzahl der Personen in den Gruppen der unterschiedlichen Ausprägung von *Anstrengungsbereitschaft in der Schule* beträgt in der Gruppe „unterdurchschnittlich“ 12, „durchschnittlich“ 18 und in der Gruppe „überdurchschnittlich“ 11; Gesamt:  $n = 41$ .

Es zeigt sich zusätzlich zum Effekt des Alters ein geringer Effekt von *Anstrengungsbereitschaft in der Schule* auf die Anzahl der wegen des Abbruchkriteriums nicht bearbeiteten Items von  $\text{Eta}^2 = 0,167$  ( $p = 0,034$ ). Die (unter Einbezug der Kovariate) geschätzten Randmittel der Anzahl nicht bearbeiteter Items liegen bei den *unterdurchschnittlich ausdauernden* im Mittel bei 17,7 bei den *durchschnittlich ausdauernden* bei 16,9 und bei den *überdurchschnittlich ausdauernden* bei 11,5 (s. Abb. 6.2). Die paarweisen Vergleiche ergeben eine Differenz von 6,2 *nicht*-bearbeiteten Items zwischen den *unterdurchschnittlich ausdauernden* und den *überdurchschnittlich ausdauernden*, sowie eine Differenz von 5,4 *nicht*-bearbeiteten Items zwischen den *durchschnittlich ausdauernden* und den *überdurchschnittlich*

*ausdauernden* ( $p = 0,092$  bzw.  $p = 0,082$  nach Anpassung für Mehrfachvergleiche nach Bonferroni).

**Abbildung 6.2:** Geschätzte Randmittel der univariaten Varianzanalyse der *Anzahl nichtbearbeiteter Items*, Faktor: *Anstrengungsbereitschaft in der Schule*, Kovariate: *Alter*



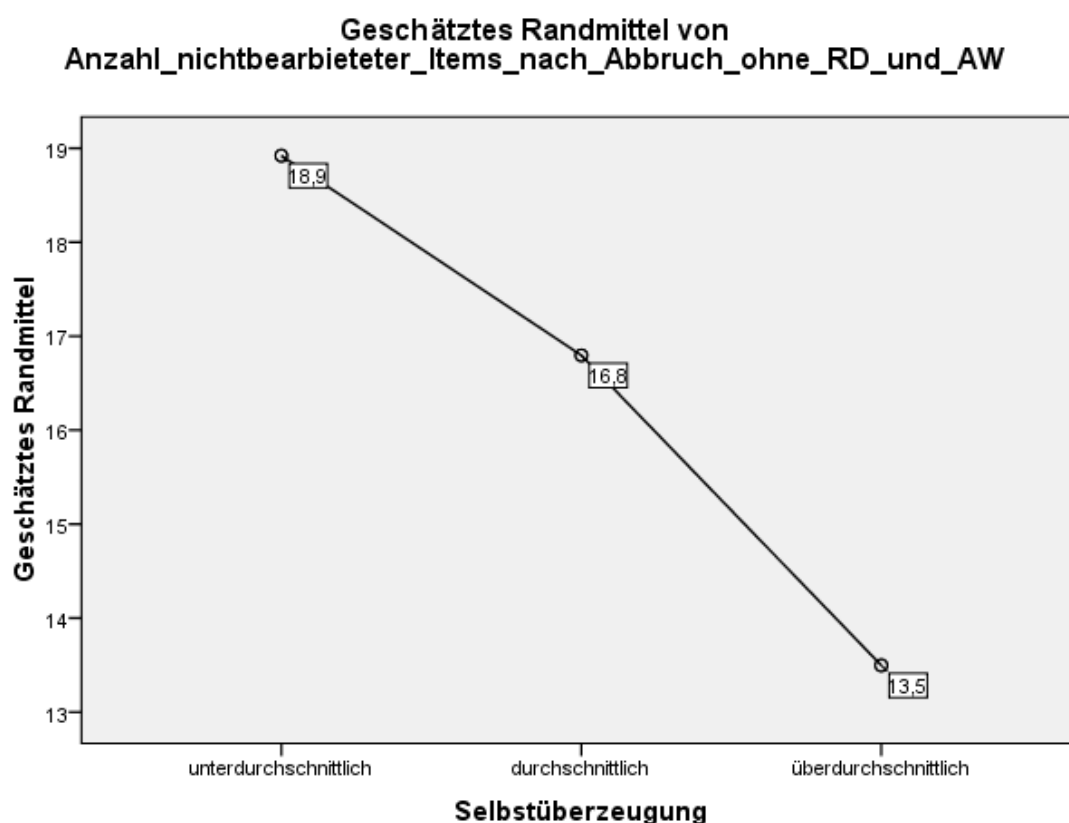
### 6.2.3 Univariate Varianzanalyse - Selbstüberzeugung

Die Anzahl der Personen in den Gruppen der unterschiedlichen Ausprägung von *Selbstüberzeugung* beträgt in der Gruppe „unterdurchschnittlich“ 6, „durchschnittlich“ 17 und in der Gruppe „überdurchschnittlich“ 18; Gesamt:  $n = 41$ .

Es zeigt sich zusätzlich zum Effekt des Alters ein nur unwesentlicher (und zufallskritisch als schlecht abgesichert) zu bewertender Effekt von *Selbstüberzeugung* auf die Anzahl der wegen des Abbruchkriteriums nicht bearbeiteten Items:  $\text{Eta}^2 = 0,10$  ( $p = 0,14$ ). Die (unter Einbezug der Kovariate) geschätzten Randmittel der Anzahl nicht bearbeiteter Items liegen bei den *unterdurchschnittlich ausdauernden* im Mittel bei 18,9 bei den *durchschnittlich ausdauernden* bei 16,8 und bei den *überdurchschnittlich ausdauernden* bei 13,5 (s. Abb. 6.3). Die paarweisen Vergleiche zeigen ein zufallskritisch nur

ungenügend abgesichertes Ergebnis ( $p = 0,22$  bzw.  $p = 0,42$  nach Anpassung für Mehrfachvergleiche nach Bonferroni). Dies zeigt eine Differenz von 5,4 *nicht*-bearbeiteten Items zwischen den *unterdurchschnittlich ausdauernden* und den *überdurchschnittlich ausdauernden*, sowie eine Differenz von 3,3 *nicht*-bearbeiteten Items zwischen den *durchschnittlich ausdauernden* und den *überdurchschnittlich ausdauernden*.

**Abbildung 6.3:** Geschätzte Randmittel der univariaten Varianzanalyse der Anzahl nichtbearbeiteter Items, Faktor: Anstrengungsbereitschaft in der Schule, Kovariate: Alter



#### 6.2.4 Zusammenfassung: durchgeführte Abbrüche und Persönlichkeitseigenschaften

Die Ergebnisse können so zusammengefasst werden, dass von den vermuteten Zusammenhängen zwischen Persönlichkeitsfaktoren und der Anzahl nichtbearbeiteter Items (als Maß für die Frühzeitigkeit der Abbrüche) auch einige empirisch gefunden werden konnten: es zeigt sich eine Tendenz, dass sehr selbstbewusste, ausdauernde, ehrgeizige, anstrengungsbereite und wenig selbstzweifelnde Testpersonen mehr Items bearbeiteten als andere, wobei diese Ergebnisse aber teilweise nur schlecht zufallskritisch abgesichert sind. Zumindest als explorative Auswertung liefert dies

damit Hinweise dafür, dass die Abbruchkriterien zu Ungleichbehandlungen von Testpersonen führen, die nicht nur auf den Zufall (der sich in der Reihenfolge der gelösten bzw. nicht gelösten Items ausdrückt), sondern systematisch auf Persönlichkeitseigenschaften zurückzuführen sind, was dem Kriterium der Fairness widerspricht.

### **6.3 Darstellungen der relativen Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)**

Bevor die *tatsächlich* durchgeführten Abbrüche nach den vorgegebenen Kriterien hinsichtlich ihrer Fairness dargestellt werden, werden folgend zuerst Grafiken zur relativen Lösungshäufigkeit der Items dargestellt. Dies geschieht, um zu zeigen, dass die Annahme der (monoton) zunehmenden Schwierigkeit der Items durch die Daten nicht eindeutig unterstützt wird. Die Untertests *Zahlen-Symbol-Test* und *Symbolsuche*, welche kein Abbruchkriterium aufweisen, sind hier genauso wenig dargestellt wie die Untertests *Zahlen Nachsprechen* und *Buchstaben-Zahlen-Folgen*, da bei diesen Untertests zum *Arbeitsgedächtnis* als gegeben erachtet werden kann, dass die Schwierigkeit der Reihen mit wachsender Anzahl der zu reproduzierenden Elemente zunimmt.

Da die Testpersonen die Items nach dem Abbruch nicht mehr bearbeiteten, sind keine Angaben vorhanden, inwiefern sie diese Items lösen hätten können. Daher werden in der Abbildung einerseits die Lösungshäufigkeiten derjenigen Testpersonen wiedergegeben, die dieses Items auch tatsächlich bearbeitet haben. Diese beschreiben gegen Ende des Untertests eine immer kleiner werdende, leistungs-selektierte Stichprobe, was zu nach oben verzerrten Werten der relativen Lösungshäufigkeiten führt. Deshalb wird andererseits auch die „abbruch-korrigierte“ Lösungshäufigkeit dargestellt: In diese werden auch die Personen miteinbezogen, die die Items (aufgrund des Abbruchkriteriums) gar nicht bearbeiteten. Die nicht bearbeiteten Items wurden als „nicht gelöst“ verrechnet, was der impliziten Annahme der Abbruchregel entspricht, die Testpersonen hätten die Items tatsächlich nicht lösen können, selbst wenn sie diese vorgegeben bekommen hätten. Die hypothetische Lösungshäufigkeit *aller* Testpersonen

bewegt sich somit innerhalb dieser Grenzen (gleich oder höher als die letztgenannte „abbruchkorrigierte“, niedriger als die nicht korrigierte).<sup>30</sup>

Für die dreikategoriell zu verrechnenden Untertests wird die dichotomisierte Form (leicht) dargestellt, da ja für das Abbruchkriterium nur relevant ist, ob das Item überhaupt gelöst wurde oder nicht (und nicht, ob es sich um 1- oder 2-Punkt-Lösungen handelt).

### 6.3.1 Mosaik-Test

**Abbildung 6.4:** *Mosaik-Test*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)

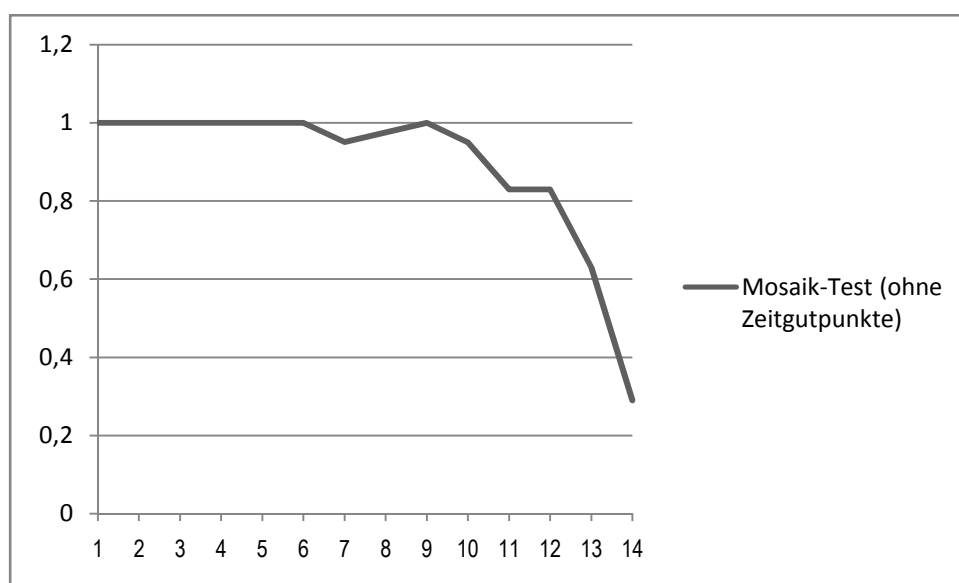


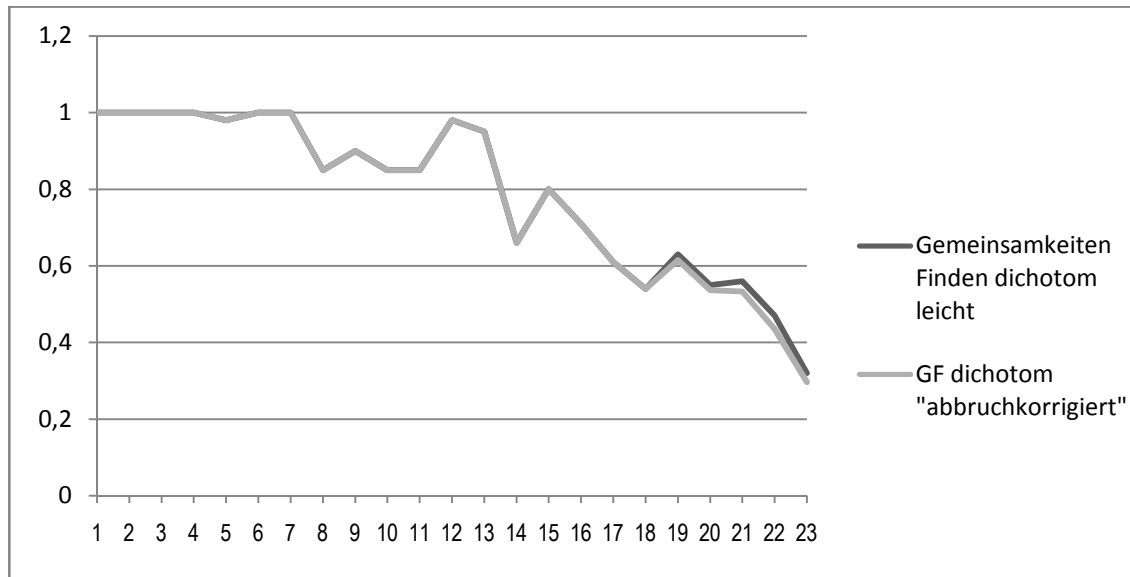
Abbildung 6.4 spricht für die Annahme der stetig steigenden Itemschwierigkeit beim *Mosaik-Test* (Abbruchkriterium: drei aufeinanderfolgend nicht gelöste Items).

---

<sup>30</sup> Theoretisch könnte die hypothetische Lösungshäufigkeit aller Testpersonen auch höher sein als die nicht korrigierte. Nämlich dann, wenn gerade jene Personen, die aufgrund des Abbruchs die nachfolgenden Items nicht vorgegeben bekamen, diese Items sogar häufiger gelöst hätten, als die Gruppe derer, bei denen das Abbruchkriterium noch nicht erfüllt war. Das wäre aber ausgesprochen unwahrscheinlich!

### 6.3.2 Gemeinsamkeiten Finden

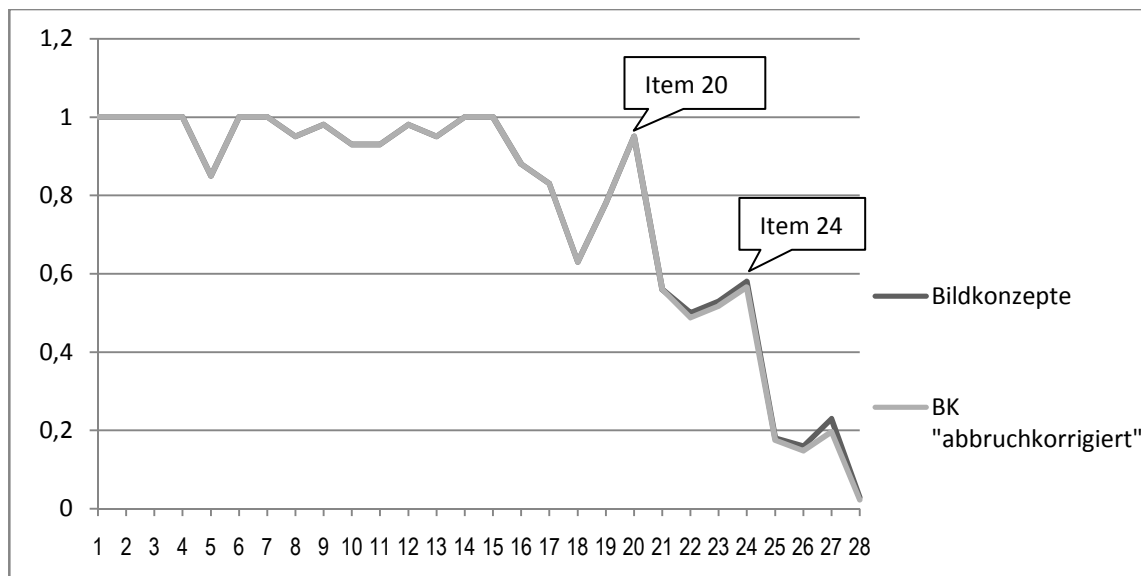
**Abbildung 6.5:** *Gemeinsamkeiten Finden*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)



Es zeigt sich in Abbildung 6.5 zum Untertest *Gemeinsamkeiten Finden* eine globale Abnahme der relativen Lösungshäufigkeiten bei den späteren Items, die aber das Abbruchkriterium dennoch nicht zweifelsfrei zu rechtfertigen scheint, da Abweichungen von der Annahme einer *stetig* steigenden Itemschwierigkeit zu erkennen sind. (Abbruchkriterium: fünf aufeinanderfolgend nicht gelöste Items)

### 6.3.3 Bildkonzepte

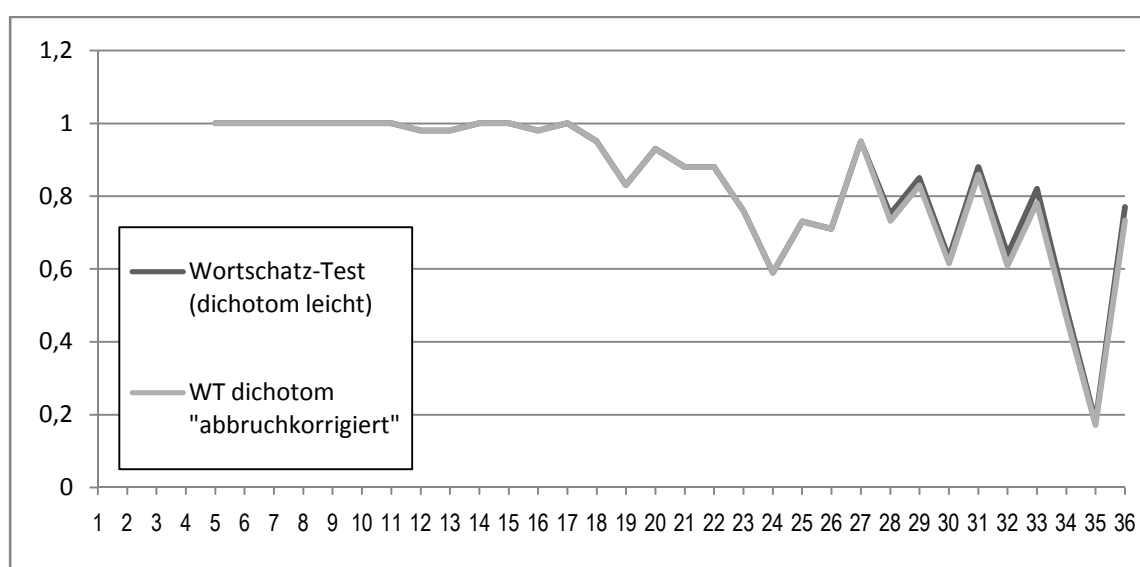
**Abbildung 6.6:** *Bildkonzepte*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)



Auch beim Untertest *Bildkonzepte* zeigen sich deutliche Abweichungen von der Annahme der stetig ansteigenden Itemschwierigkeit (s. Abb. 6.6): So wäre etwa ein Abbruch vor dem Item 20 oder dem Item 24 mit der Begründung, das nachfolgende Item wäre für die Testperson praktisch unlösbar, kaum gerechtfertigt, da diese eine hohe Lösungshäufigkeit aufweisen, als die unmittelbar vorgereichten Items (Abbruchkriterium: fünf aufeinanderfolgend nicht gelöste Items).

#### 6.3.4 Wortschatz-Test

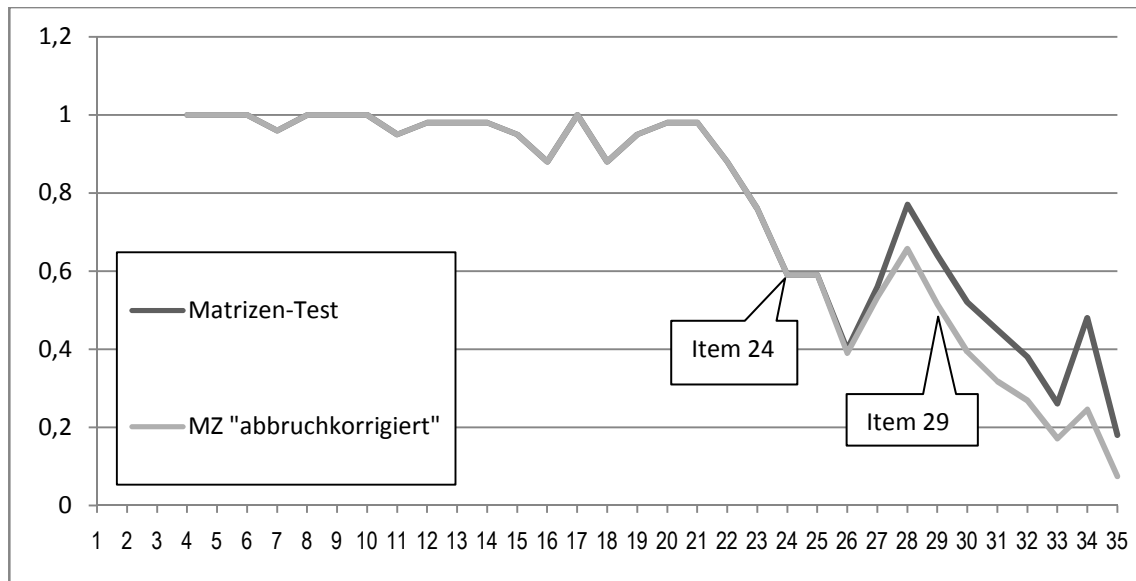
**Abbildung 6.7:** *Wortschatz-Test*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)



Nach Abbildung 6.7 nehmen die relativen Lösungshäufigkeiten des *Wortschatz-Tests* im Bereich zwischen 23 und 33 kaum ab, was jedenfalls gegen die Durchführung der Abbruchregel spricht (Abbruchkriterium: fünf aufeinanderfolgend nicht gelöste Items).

### 6.3.5 Matrizen-Test

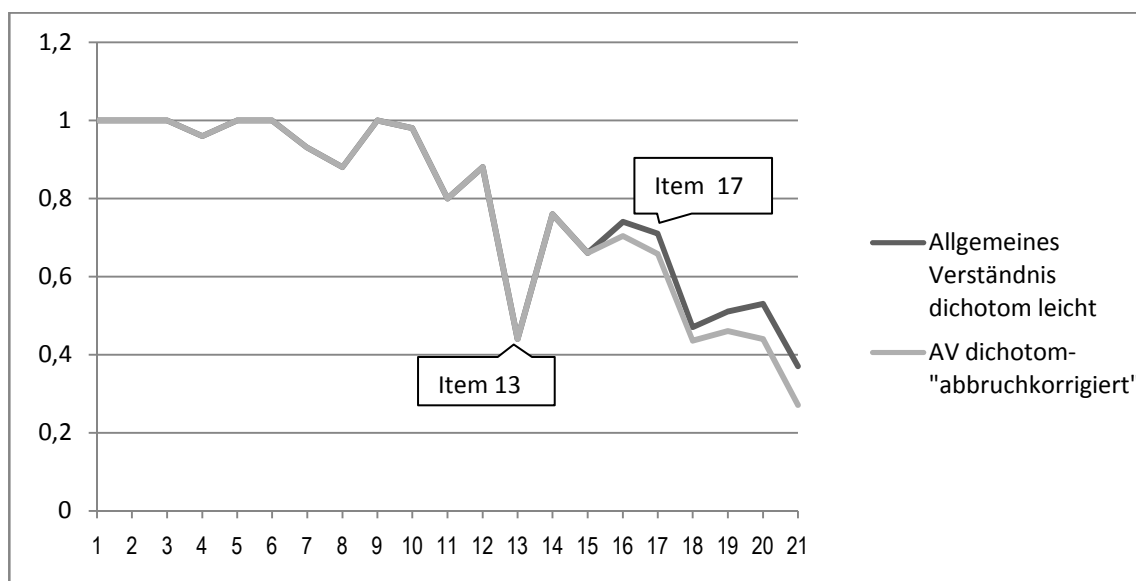
**Abbildung 6.8:** *Matrizen-Test*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)



Zwar zeigt sich im *Matrizen-Test* (Abbruchkriterium: vier Items hintereinander nicht gelöst oder vier von fünf Items nicht gelöst) eine globale Abnahme der relativen Lösungshäufigkeiten (s. Abb. 6.8), dennoch ist diese gerade im Bereich zwischen den Items 24 und 29 kaum feststellbar.

### 6.3.6 Allgemeines Verständnis

**Abbildung 6.9:** *Allgemeines Verständnis*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)

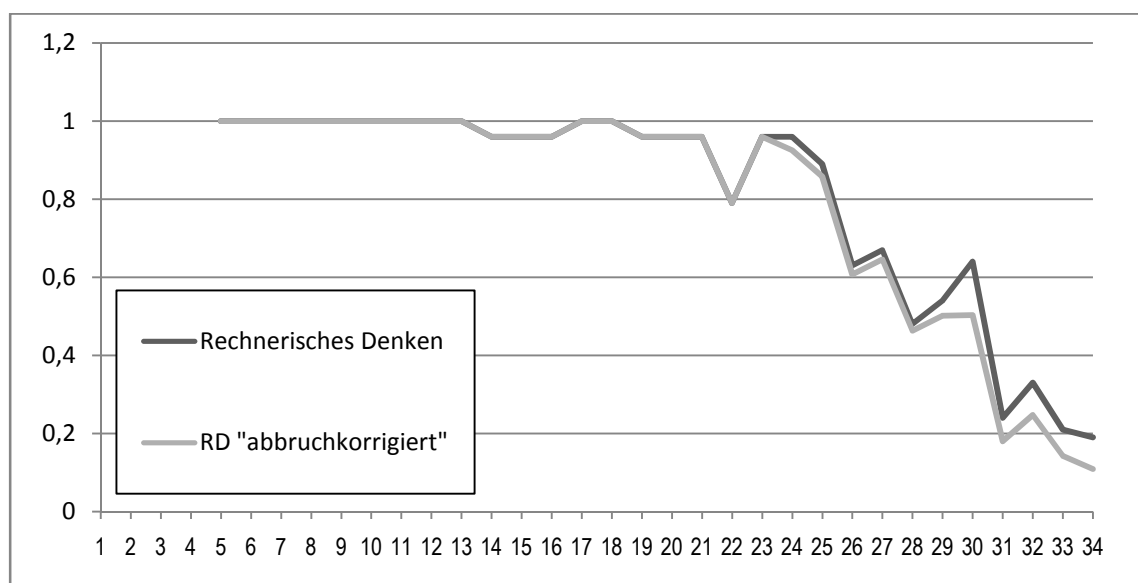




Im Untertest *Allgemeines Verständnis* (s. Abb. 6.9) zeigt sich beispielsweise im Bereich der Items 13 bis 17 eher ein Anstieg der Lösungshäufigkeit als ein Abfall, was deutlich gegen die Annahme spricht, weitere Items nach dem je individuellen Abbruch wären ohnehin praktisch „unlösbar“ (Abbruchkriterium: nach vier aufeinanderfolgend nicht gelösten Items).

### 6.3.7 Rechnerisches Denken

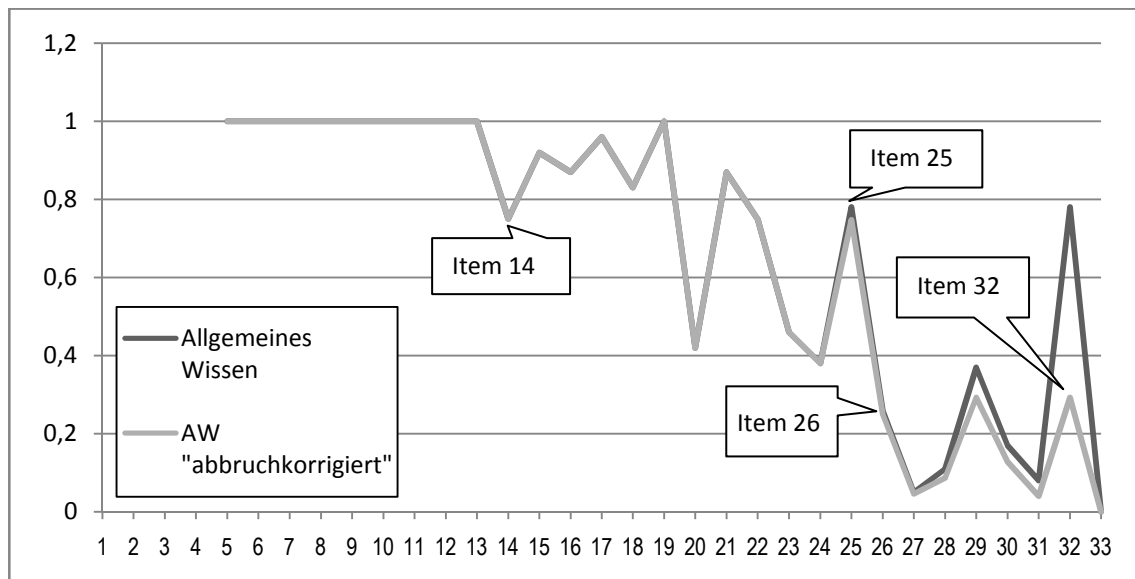
**Abbildung 6.10:** *Rechnerisches Denken*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)



Der Untertest *Rechnerisches Denken* (Abbruchkriterium: vier aufeinanderfolgend nicht gelöste Items) scheint im Wesentlichen der Annahme der stetigen Zunahme der Itemschwierigkeit zu entsprechen (s. Abb. 6.10).

### 6.3.8 Allgemeines Wissen

**Abbildung 6.11:** *Allgemeines Wissen*: relative Lösungshäufigkeiten der Items (im Sinne des Abbruchkriteriums)



Beim Untertest *Allgemeines Wissen* (Abb. 6.11) ist sowohl im Bereich zwischen den Items 14 und 25, als auch im Bereich zwischen den Items 26 und 32 keine stetige Zunahme der Itemschwierigkeit feststellbar, was klar gegen die Anwendung der Abbruchregel spricht. (Abbruchkriterium: fünf aufeinanderfolgend nicht gelöste Items)

## 6.4 Darstellung der durchgeführten Abbrüche hinsichtlich ihrer Fairness

Es wurde schon erwähnt, dass die Abschätzung der Leistungsfähigkeit einer Testperson, wie sie im Rahmen der Abbruchkriterien erfolgt, in vielen Fällen im Widerspruch zu der Abschätzung der Leistung steht, wie sie zur Bestimmung der entsprechenden Wertpunkte bzw. IQ-Punkte vorgenommen wird. Dies zieht den Schluss nach sich, dass (zumindest) eine der beiden Vorgangsweisen nicht *verrechnungsfair* sein kann. Dass die Verrechnung der Gesamtzahl gelöster Items (oder erreichter Punkte) eigentlich einer Überprüfung durch ein probabilistisches Modell bedarf, um als *verrechnungsfair* bewertet zu werden, wurde unter Punkt 6.1 erwähnt. Sie steht also im Falle des HAWIK-IV in Ermangelung dieser Überprüfung sozusagen „auf tönernen Füßen“. Bei Betrachtung der Verrechnungsvorschrift für die Abbruchkriterien lässt sich dagegen sagen: Die *alleinige* Verrechnung nur der jeweils *letzten* drei bis fünf Items *ohne* genaue Bestimmung der Schwierigkeitsparameter dieser Items steht (um bei dem Bild zu bleiben) auch auf „tönernen Füßen“, aber außerdem auch nur auf ganz *wenigen* dieser zerbrechlichen Stützen. Sie scheint daher noch weniger zuverlässig zu sein. Aus diesem Grund werden

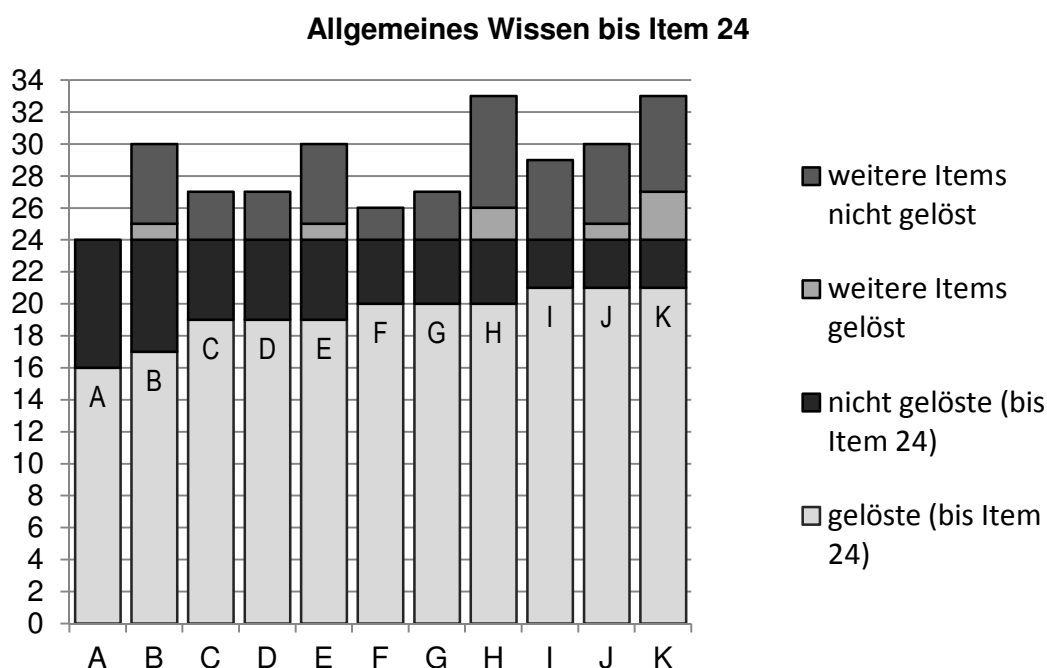
Widersprüche zwischen diesen Verrechnungsvorschriften als klare Kritik an der letztgenannten Verrechnungsart (also an den Abbruchkriterien) gewertet. In Zusammenhang mit den bisher dargestellten Ergebnissen betreffend Persönlichkeitseigenschaften einerseits und relativen Lösungshäufigkeiten andererseits werden diese Widersprüche als Indiz für die mangelnde Fairness der Abbruchkriterien gewertet.

Zur Darstellung der tatsächlich durchgeführten Abbrüche (die auf mangelnde Fairness der Untertests schließen lassen) werden für alle Untertests Grafiken erstellt. Diese Abbildungen sind sozusagen „Momentaufnahmen“ aller Testpersonen nach einem bestimmten Item (beispielsweise nach dem Item, das noch alle Testpersonen bearbeiteten). Die Untertests sind nach ihrer Zugehörigkeit zu den Indizes geordnet. Der Aufbau der Grafiken wird hier anhand des Untertests *Allgemeines Wissen* erklärt.

#### 6.4.1 Abbruchregel Allgemeines Wissen

Beim Untertest *Allgemeines Wissen* hat der Abbruch nach fünf aufeinanderfolgenden nicht gelösten Items zu erfolgen.

**Abbildung 6.12:** Abbrüche *Allgemeines Wissen*: gelöste und nicht gelöste Items **bis zum** Item 24 und **nach dem** Item 24 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 24“ leistungsschwächsten elf Testpersonen



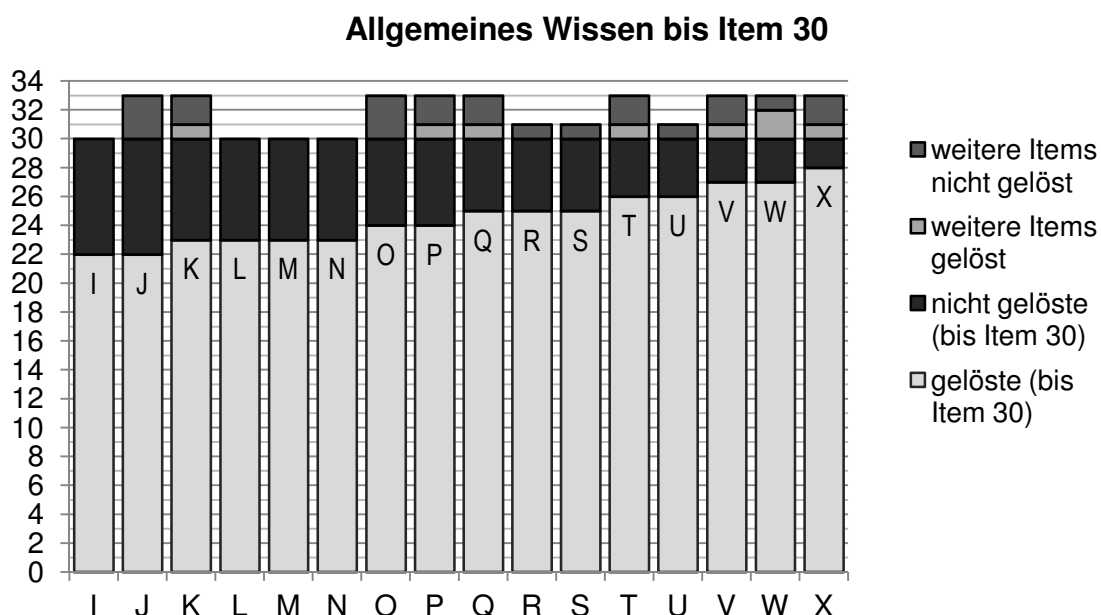
Die einzelnen Säulen der Abbildung 6.12 repräsentieren die einzelnen Testpersonen, wobei diese nach ihren erreichten Punkten bis zum Item 24 gerangreicht wurden. Zu sehen sind nur die elf Testpersonen mit den wenigsten gelösten Items, da die anderen Testpersonen noch deutlich mehr Items bearbeiten konnten und somit für die Situation nach 24 Items keine Informationen liefern; sie werden erst in einer späteren Grafik dargestellt. Die Rangplätze der Testpersonen sind, nicht durch Zahlenwerte, sondern durch Buchstaben von A bis Z, gefolgt von AA, BB, CC... angegeben.

Die untersten Teile der Säulen repräsentieren jene Items, welche (bis zum Item 24) als gelöst verrechnet wurden (inklusive der Items, die aufgrund der Einstiegsregeln zwar nicht bearbeitet, aber dennoch als gelöst verrechnet wurden); die darauf liegenden Säulenteile repräsentieren die (bis zum Item 24) bearbeiteten, aber nicht gelösten Items.

Die Summe dieser zwei Kategorien ergibt demnach bei allen Testpersonen 24. Über dem Wert 24 ist nun für alle Testpersonen eingezeichnet, wie viele Items sie nach dem Item 24 noch zur Bearbeitung bekamen: zuerst die, die sie lösen konnten und darüber die, die sie nicht lösen konnten. Testperson A beispielsweise hat bis zum 24. Item 16 Items gelöst (bzw. acht Items nicht gelöst) und dabei das Abbruchkriterium bereits erfüllt. Testperson B hat zwar auch sieben Items nicht lösen können, allerdings in einer solchen Reihenfolge, dass das Abbruchkriterium nicht erfüllt wurde. Obwohl sie nur ein Item mehr lösen konnte als Person A bekommt sie um sechs Items mehr zur Bearbeitung!

Betrachten wir nun die Testpersonen mit den Rangplätzen B bis E, so erkennen wir, dass die Testpersonen C und D nach dem Item 24 nur mehr drei Items vorgelegt bekommen, während die Testpersonen B und E noch sechs Items bearbeiten dürfen (und z.T. auch lösen), obwohl diese bis dahin entweder gleich viele oder sogar weniger Items gelöst haben als die Testpersonen C und D. Auch die Personen mit den Rangplätzen F und G sind gegenüber den Personen B, E und H klar benachteiligt!

**Abbildung 6.13:** Abbrüche *Allgemeines Wissen*: gelöste und nicht gelöste Items **bis zum Item 30** und **nach dem Item 30** vorgegebene gelöste bzw. nicht gelöste Items; die leistungsschwächsten acht Testpersonen sind nicht dargestellt



In Abbildung 6.13 desselben Untertests nach dem Item 30 sind nur die Testpersonen ab dem Rangplatz 9 (= I) dargestellt. (Die Rangplätze werden für jede Abbildung gesondert berechnet, sind also unterschiedlich zur vorigen Grafik!): obwohl die Testperson J bis dahin gleich viele Items gelöst hat wie Person I und sogar weniger Items gelöst hat, als alle Personen mit höheren Rängen, bekommt sie *mehr* weitere Items zur Bearbeitung als die Personen I, L, M, N, R, S und U und ist deshalb diesen gegenüber bevorzugt!<sup>31</sup> Auch Person K darf – trotz gleicher Anzahl gelöster Items – um drei Items mehr bearbeiten als die Personen L, M und N, die das Abbruchkriterium (aufgrund einer anderen Reihenfolge der gelösten bzw. nicht gelösten Items) bereits erfüllen!

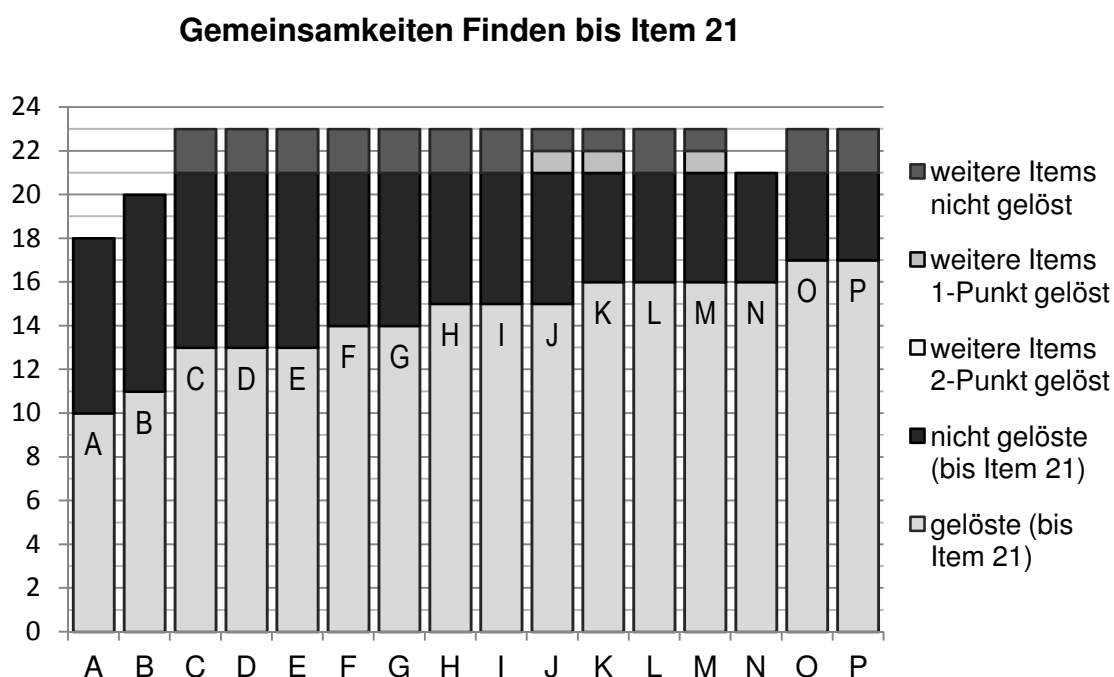
Es kann zusammengefasst werden, dass selbst bei gleicher Anzahl bisher gelöster Items die Abbruchregel zu Unterschieden von mehreren Items führt, die bei einer Person noch vorgegeben werden, bei einer anderen jedoch nicht, was gegen das Kriterium der Fairness spricht.

<sup>31</sup> Einer gelösten zusätzlichen Aufgabe entspricht in den allermeisten Altersgruppen und Fähigkeitsbereichen etwa ein Wertpunkt!

#### 6.4.2 Abbruchregel Gemeinsamkeiten Finden

Bei diesem Untertest ist abzubrechen, wenn die Testperson fünf aufeinanderfolgende Aufgaben nicht löst.

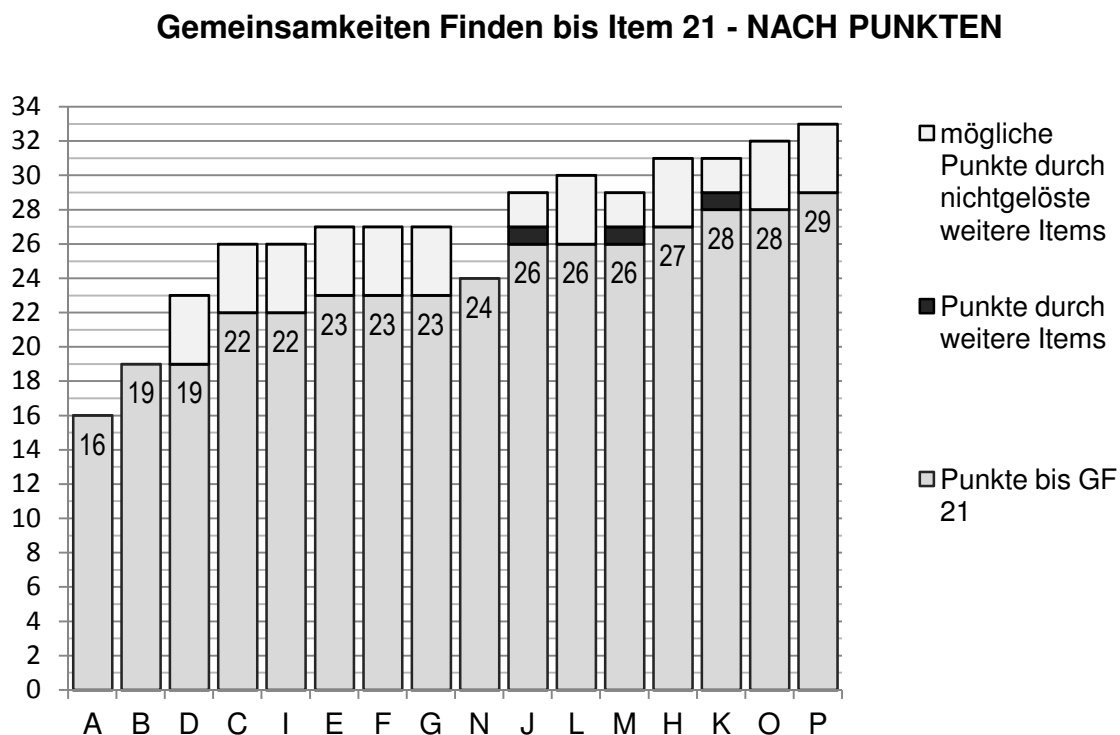
**Abbildung 6.14:** Abbrüche *Gemeinsamkeiten Finden*: gelöste und nicht gelöste Items **bis zum** Item 21 und **nach dem** Item 21 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 21“ leistungsschwächsten 16 Testpersonen



Dargestellt ist hier die Situation nach 21 vorgegebenen Items (Abb. 6.14). Man sieht, dass nur drei Testabbrüche vorkommen, wobei allerdings einer davon schon klar aufzeigt, dass auch dieser Untertest nicht vor ungerechtfertigten Abbrüchen bewahrt bleibt. Testperson N hat zwar gleich viele oder mehr Items gelöst wie die Testpersonen C bis M, aber eben das „Pech“, dass alle fünf nicht gelösten Items unmittelbar aufeinanderfolgen, was dazu führt, dass sie keine weiteren Items mehr vorgelegt bekommt, während die anderen Personen noch zwei weitere Items bearbeiten dürfen.

Da die Lösungen dieses Untertests mit ein oder zwei Punkten bewertet werden, ist die Anzahl der bisher (überhaupt) gelösten Items keine hinreichende Darstellung der bisherigen Leistung, weshalb hier auch eine weitere Abbildung (6.15) dargestellt ist, bei der nicht die Anzahl der gelösten Items dargestellt wird, sondern die dadurch erworbenen Punkte.

**Abbildung 6.15:** Abbrüche *Gemeinsamkeiten Finden*: **bis zum** Item 21 erreichte Punkte und **nach dem** Item 21 erreichte Punkte bzw. *mögliche* Punkte durch nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 21“ leistungsschwächsten 16 Testpersonen



Die Anordnung der Testpersonen auf der x-Achse in dieser Grafik richtet sich nach der Anzahl der bis zum Item 21 erworbenen Punkte, die Bezeichnungen mit Buchstaben entspricht der der oberen Grafik. Die untersten Teile der Säulen entsprechen den bisher erlangten Punkten, die obersten Teile entsprechen den potentiellen Punkten, die durch die weiteren Items, die *nicht* gelöst wurden, erlangt werden hätten können. Die dazwischenliegenden Säulenteile (zu sehen bei den Personen J, M, K) zeigen jene Punkte, die durch weitere gelöste Items erworben werden.

Die vorher beschriebene Testperson N hat 24 Punkte und ist dadurch gekennzeichnet, dass ihr kein weiteres Item vorgelegt wird, während die sechs (links davon dargestellten) Testpersonen (mit Punkten zwischen 19 und 23) jeweils noch vier (!) weitere Items bearbeiten dürfen. Dass es diesen Testpersonen nicht gelingt, diese Items zu lösen, rechtfertigt dennoch nicht, dass Person N von dieser Chance ausgeschlossen wird! <sup>32</sup>

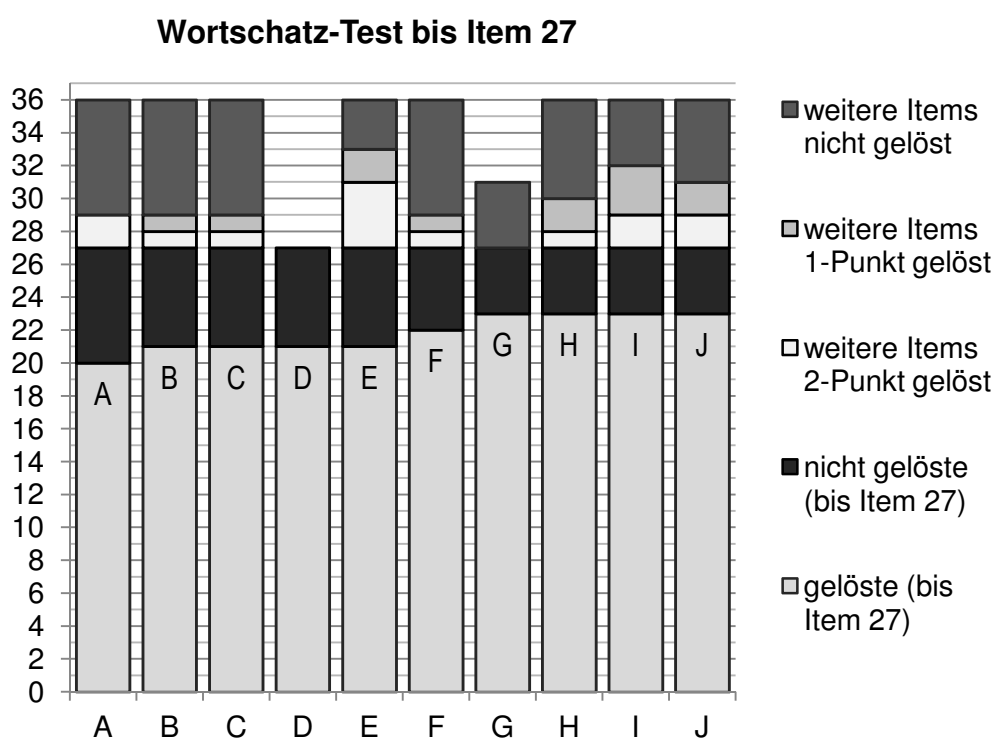
<sup>32</sup> Je nach Altersgruppe entsprechen einem Wertpunkt 2 – 3 zusätzliche Punkte, d.h. schon das mit zwei Punkten bewertete Lösen eines zusätzlichen Items führt in den meisten Fällen zu einer Zunahme der Wertpunkte!

Es lässt sich zusammenfassen, dass trotz der geringen Anzahl an Abbrüchen zumindest *ein* Abbruch zu erkennen ist, der nicht durch die gesamte bisherige Leistung in diesem Untertest begründet ist und damit ungerechtfertigt erscheint.

#### 6.4.3 Abbruchregel Wortschatztest

Bei diesem Untertest ist abzubrechen, wenn die Testperson fünf aufeinanderfolgende Aufgaben nicht löst.

**Abbildung 6.16:** Abbrüche *Wortschatz-Test*: gelöste und nicht gelöste Items **bis zum** Item 27 und **nach dem** Item 27 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 27“ leistungsschwächsten 10 Testpersonen

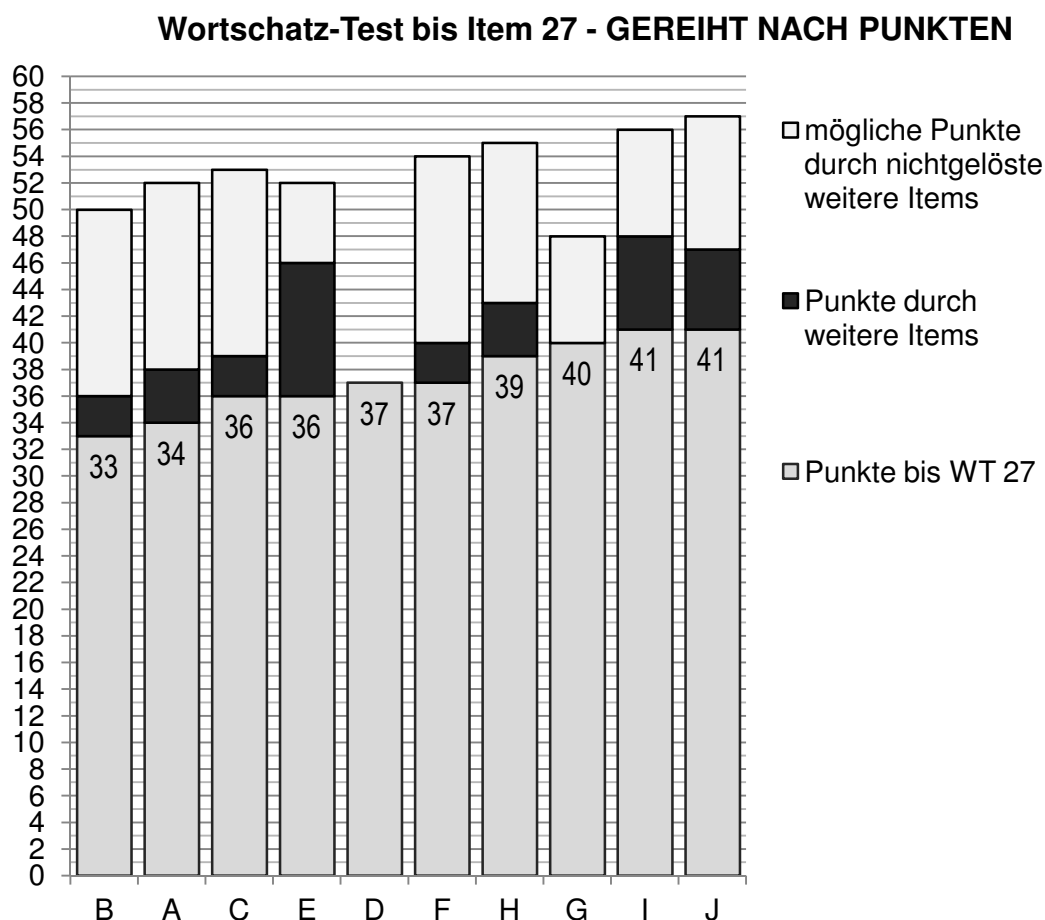


Dargestellt ist in Abbildung 6.16 die Situation nach 27 vorgegebenen Items. Es sind zwar nur zwei Testpersonen zu sehen, bei denen die Abbruchregel zur Anwendung kam, doch wurden der einen Testperson (G) fünf Items weniger vorgegeben und der anderen Testperson (D) sogar neun Items weniger als allen anderen Testpersonen!

Da die Lösungen dieses Untertests mit ein oder zwei Punkten bewertet werden, wird auch hier eine andere Grafik dargestellt (Abb. 6.17), bei der nicht die Anzahl der gelösten Items dargestellt wurde sondern die dadurch erworbenen Punkte. Die Anordnung der Testpersonen auf der x-Achse in dieser Grafik richtet sich nach den bis zum Item 27 erworbenen Punkten.



**Abbildung 6.17:** Abbrüche *Wortschatz-Test*: **bis zum** Item 27 erreichte Punkte und **nach dem** Item 27 erreichte Punkte bzw. *mögliche* Punkte durch nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 27“ leistungsschwächsten 10 Testpersonen



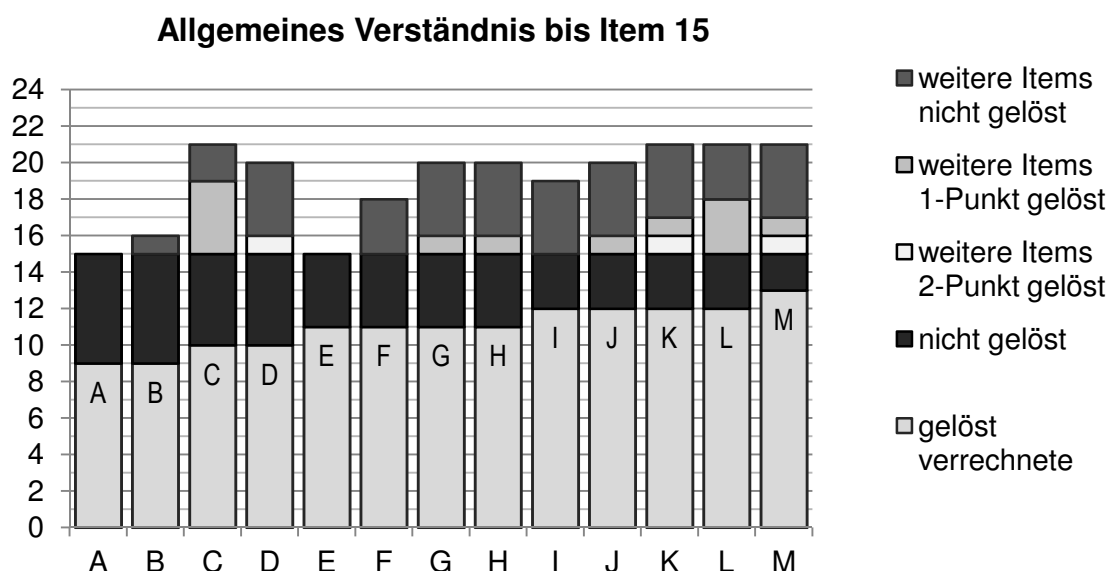
Wiederum ist zu erkennen, dass dieselben zwei Personen (D und G) gegenüber anderen Testpersonen mit weniger oder gleich vielen Punkten benachteiligt werden. Auffällig ist auch, dass die anderen Testpersonen mit bis zu 37 Punkten (A, B, C, E, F) noch mindestens drei Punkte durch die zusätzlich vorgegebenen Items erhielten, eine Testperson (E) sogar zehn Punkte!<sup>33</sup> Zwar zeigen sich in diesem Untertest nur zwei benachteiligte Personen, was auch am bereits beschriebenen Deckeneffekt liegt (d.h. es gibt zu wenig schwierige Items, weswegen auch nur wenige Testpersonen das Abbruchkriterium erfüllen), diese sind aber zumindest in *einem* Fall durchaus schwerwiegend!

<sup>33</sup> Im Alters- und Fähigkeitsbereich der Testperson D entsprechen drei Punkten etwa ein Wertpunkt. Es ist dies mit 11 Wertpunkten ihre schlechteste Untertestleistung im Index Sprachverständnis, in dem sie den Durchschnittswert von 13,5 Wertpunkten erreicht.

#### 6.4.4 Abbruchregel Allgemeines Verständnis

Bei diesem Untertest ist abzubrechen, wenn die Testperson vier aufeinanderfolgende Aufgaben nicht löst. Dargestellt ist hier die Situation nach 15 vorgegebenen Items.

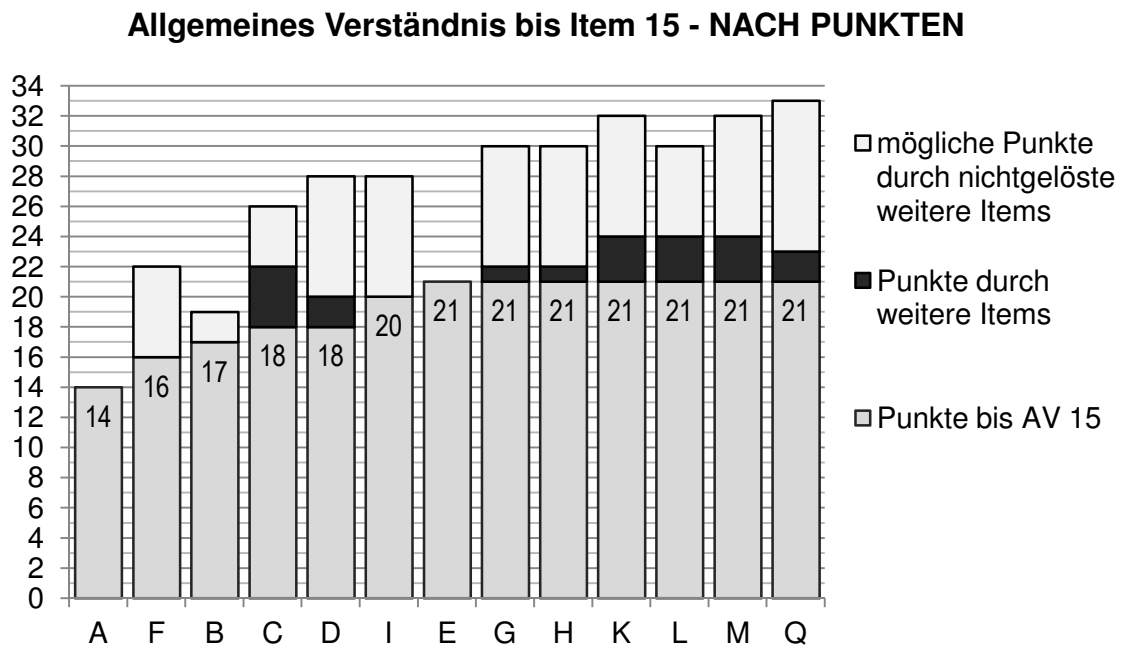
**Abbildung 6.18:** Abbrüche *Wortschatz-Test*: gelöste und nicht gelöste Items **bis zum** Item 15 und **nach** dem Item 15 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 15“ leistungsschwächsten 13 Testpersonen



In Abbildung 6.18 zeigt sich (zumindest) eine von der Abbruchregel deutlich benachteiligte Testpersonen, was auch anhand der unteren Grafik (Abb. 6.19) deutlich wird, die auf den bisher erreichten Punkten und nicht auf der Anzahl überhaupt gelöster Items basiert. Diese Testperson (E) ist in Abbildung 6.19 mit der Punkteanzahl 21 zu sehen. Im Vergleich sowohl zu den Testpersonen mit der gleichen (oder niedrigeren) Anzahl bisher gelöster Items (das sind die Testpersonen B, C, D, F, G, H) als auch im Vergleich zu den Testpersonen mit der gleichen (oder niedrigeren) erreichten Punkteanzahl (das sind sogar elf Testpersonen!) bekommt diese Testperson um bis zu sechs Items weniger zur Bearbeitung!<sup>34</sup>

<sup>34</sup> In der Altersgruppe und dem Fähigkeitsbereich dieser Testperson entsprechen einem Wertpunkt ein bis zwei Punkte; das heißt, schon mit einer einzigen zusätzlichen 2-Punkt-Antwort und einer Ein-Punkt-Antwort wären für diese Testperson zwei weitere Wertpunkte zu erlangen gewesen.

**Abbildung 6.19:** Abbrüche *Allgemeines Verständnis*: **bis zum** Item 15 erreichte Punkte und **nach dem** Item 15 erreichte Punkte bzw. *mögliche* Punkte durch nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 15“ leistungsschwächsten 13 Testpersonen



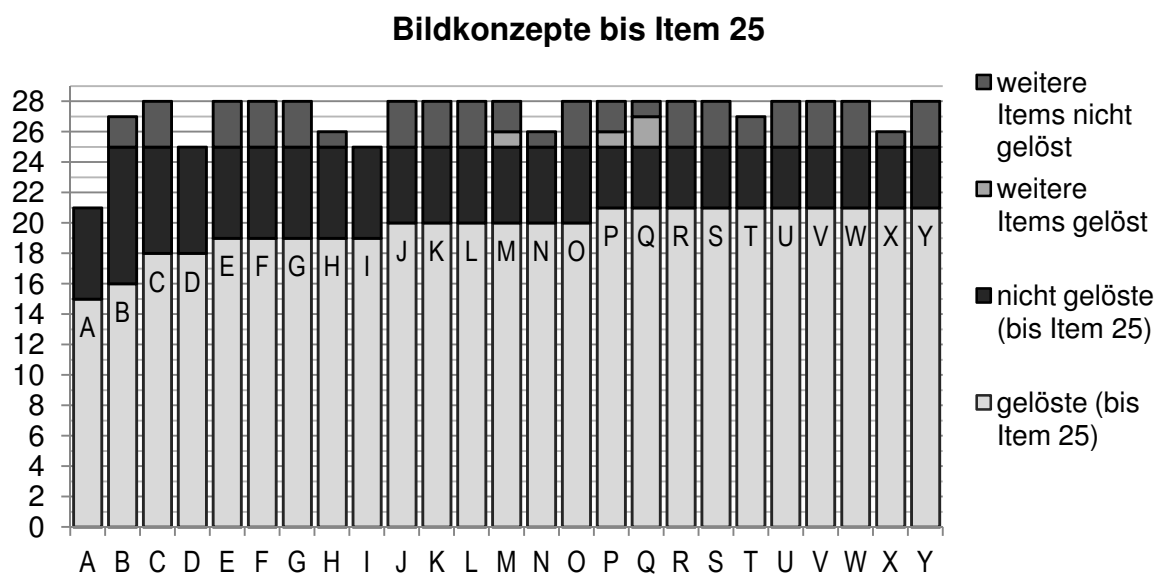
#### 6.4.5 Abbruchregel *Mosaik-Test*

Beim *Mosaik-Test* ist abzuberechnen, wenn die Testperson drei aufeinanderfolgende Aufgaben nicht löst. Da in der Stichprobe dieser Arbeit überhaupt nur drei Abbrüche vorkamen und diese nur die letzten beiden Items betrafen, wurden keine Hinweise dafür gefunden, dass die Abbruchregel zu Verstößen gegen das Kriterium der Fairness führt.

#### 6.4.6 Abbruchregel Bildkonzepte

Bei diesem Untertest ist abzubrechen, wenn eine Testperson fünf aufeinanderfolgende Aufgaben nicht löst.

**Abbildung 6.20:** Abbrüche *Bildkonzepte*: gelöste und nicht gelöste Items **bis zum** Item 25 und **nach dem** Item 25 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 25“ leistungsschwächsten 25 Testpersonen



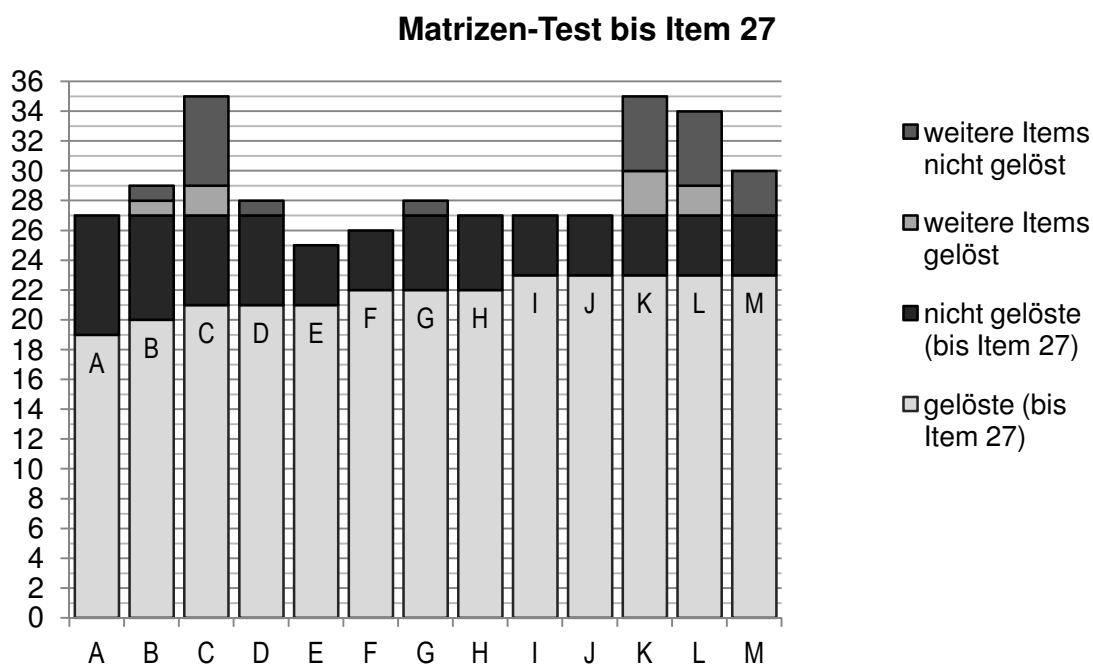
Dargestellt ist hier die Situation nach 25 vorgegebenen Items (Abb. 6.20). Es sind einige Testpersonen zu erkennen, die bis zu drei Items weniger zur Bearbeitung bekommen (z.B. D, H, I, N) als andere Testpersonen, die genauso viele oder weniger bisherige Items lösen konnten, was durchaus zu Benachteiligungen führen kann, auch wenn nur wenige der Testpersonen die zusätzlichen Items wirklich lösen können.<sup>35</sup>

#### 6.4.7 Abbruchregel Matrizen-Test

Bei diesem Multiple-Choice-Test sind Matrizenaufgaben zu lösen, wobei eine von fünf Möglichkeiten die Lösung darstellt. Somit hat jedes Item also eine Ratewahrscheinlichkeit von 20 %. Abbruchkriterium beim Matrizentest sind vier nicht gelöste Items innerhalb von fünf aufeinanderfolgenden Items.

<sup>35</sup> Eine gelöste zusätzliche Aufgabe führt beim Untertest *Bildkonzepte* in den allermeisten Altersgruppen und Fähigkeitsbereichen zu einer Zunahme von einem Wertpunkt.

**Abbildung 6.21:** Abbrüche *Matrizen-Test*: gelöste und nicht gelöste Items **bis zum Item 27** und **nach dem Item 27** vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 27“ leistungsschwächsten 13 Testpersonen

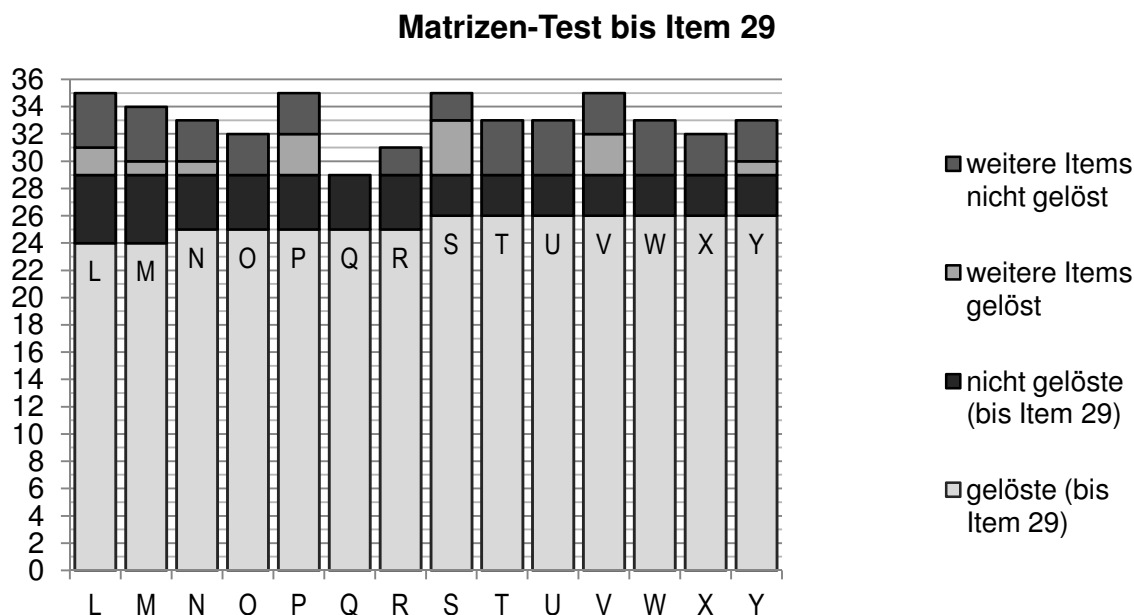


Die erste Grafik (Abb. 6.21) stellt die Situation nach 27 vorgelegten Items dar. Auch in diesem Untertest zeigen sich einige Unterschiede in der Anzahl der weiteren vorgegebenen Items, die nicht durch die bisherige Gesamtleistung begründet sind, was zu erheblichen Verstößen gegen die Fairness führt: So hat etwa die Testperson C (nach 27 Items) zwar sechs Items nicht gelöst, bekommt aber dennoch acht weitere Items vorgelegt (von denen sie zwei löst, und sechs nicht), während die Personen E, F, I und J nur vier nicht gelöst haben, aber diese eben in einer solchen Reihenfolge, dass das Abbruchkriterium – und man beachte dabei den Zufallsaspekt – bereits erfüllt wurde!

Genauso wenig durch die Anzahl der bisher gelösten Items begründet ist der massive Unterschied zwischen den Personen I und J einerseits und den Personen K, L und M andererseits. Alle fünf Personen konnten bis zum Item 27 nur vier Items nicht lösen. Die Personen I und J bekommen keine weiteren Items, die anderen drei Personen noch drei bis acht Items, die sie z.T. auch lösen können.<sup>36</sup>

<sup>36</sup> Je nach Altersgruppe und Fähigkeitsbereich entsprechen einem Wertpunkt 1-2 gelöste Items.

**Abbildung 6.22:** Abbrüche *Matrizen-Test*: gelöste und nicht gelöste Items **bis zum** Item 29 und **nach dem** Item 29 vorgegebene gelöste bzw. nicht gelöste Items; die zum Zeitpunkt „Item 29“ leistungsschwächsten 11 und leistungstärksten 16 Testpersonen sind nicht dargestellt



Bei Betrachtung von Abbildung 6.22 zum Matrizentest nach dem Item 29 (neue Rangreihe und Buchstabenzuordnung) zeigt sich zumindest *ein* weiterer deutlich benachteiligender, nicht durch die Anzahl der bisher gelösten Aufgaben gerechtfertigter Abbruch: Person Q darf, trotz gleich guter bzw. besserer Leistung wie die Personen L bis R keine weiteren Items bearbeiten, während die anderen genannten Personen noch zwei bis sechs weitere Items bearbeiten dürfen, wobei der einzige Unterschied in der Reihenfolge der gelösten bzw. nicht-gelösten Items besteht!

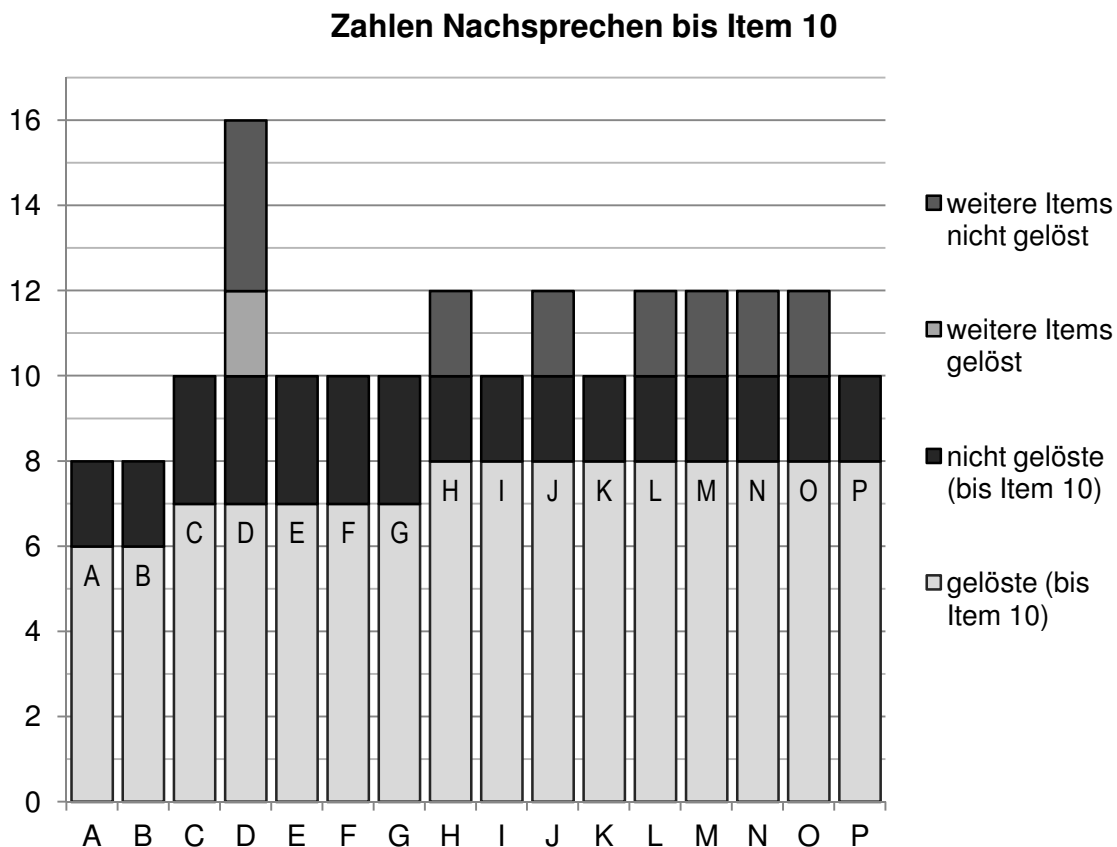
Im Falle des *Matrizen-Tests* ist als Besonderheit zu bedenken, dass sogar das rein zufällige Erraten der Lösung eines Items, was zu einem Verhindern des Abbruchs führt, recht wahrscheinlich ist: Die Abbruchregel stützt sich ja auf die Argumentation, dass die Items streng monoton nach ihrer Schwierigkeit geordnet wären und deshalb anzunehmen wäre, dass wer die letzten paar Items nicht lösen konnte, die weiteren (ja demnach deutlich schwierigeren) Items auch nicht wird lösen können. Abgesehen davon, dass die Annahme der monoton zunehmenden Itemschwierigkeit nicht zuzutreffen scheint, wird im Falle des *Matrizen-Tests* folgendes offensichtlich: Sogar ein *rein zufälliges* Lösen von Items und damit das Verhindern eines Abbruchs ist sehr wahrscheinlich, da die Lösung jedes Items nur aus fünf Möglichkeiten herausgesucht werden muss. Es besteht also eine a-priori-Ratewahrscheinlichkeit von 1/5! Nach der Binomialverteilung hat ein rein zufälliges Lösen von zumindest einem Item innerhalb

von vier Items (und damit das Verhindern des Abbruchs) sogar die Wahrscheinlichkeit von 0,59! Im Falle des *Matrizen-Tests* ist also die mangelnde Fairness der Abbruchregel nicht nur durch die Daten zu belegen, sondern aufgrund des Rateeffekts fast zwingend zu erwarten!

#### 6.4.8 Abbruchregel Zahlen Nachsprechen

Der Untertest *Zahlen Nachsprechen* besteht aus zwei Teilen, dem ZN *vorwärts* und dem ZN *rückwärts*, deren Scores addiert werden, um diese Summe in Wertpunkte umzuwandeln. Bei diesem Untertest werden der Testperson Zahlenreihen vorgesagt, die sie nachsprechen soll (*vorwärts*: in der vorgedachten Reihenfolge, *rückwärts*: in der umgekehrten Reihenfolge), wobei jeweils zwei Zahlenreihen mit gleicher Länge (von zwei bis neun Elementen) vorgegeben werden; der Untertest ist abzubrechen, wenn die Testperson *beide* Zahlenreihen mit *der selben* Anzahl von Zahlen nicht fehlerfrei nachsprechen kann.

**Abbildung 6.23:** Abbrüche *Zahlen Nachsprechen vorwärts*: gelöste und nicht gelöste Items **bis zum** Item 10 und **nach dem** Item 10 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 10“ leistungsschwächsten 16 Testpersonen



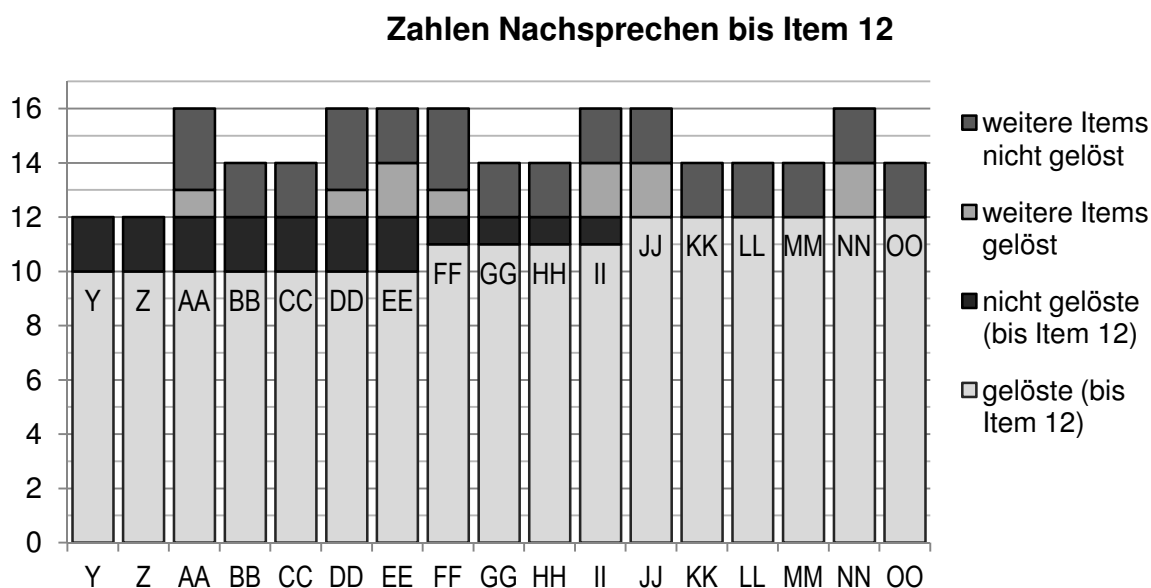
*Zahlen Nachsprechen vorwärts:* Dargestellt ist zuerst die Situation nach zehn vorgegebenen Items, d.h. nach beiden Aufgaben mit jeweils sechs Zahlen (Abb. 6.23). Auf den ersten Blick ist zu sehen, dass die Testperson D genauso viele Zahlenreihen nachsprechen konnte, wie die Testpersonen C bis G und sogar weniger als die Testpersonen H bis P (und noch weitere!), dennoch werden ihr noch *sechs* Zahlenreihen vorgegeben, während von den anderen oben genannten sieben Testpersonen *kein einziges* weiteres Items bearbeiten dürfen und sechs Testpersonen nur *zwei weitere* Items (die weiteren Personen sind hier nicht dargestellt). Testperson D ist den anderen Personen gegenüber bevorzugt, in dem Sinne, dass sie ihr Leistungsmaximum zeigen kann, obwohl ihr dazwischen auch Fehler unterlaufen, während andere wegen der gleichen Anzahl von Fehlern einen Testabbruch erfahren müssen.

Dagegen erscheinen die Testpersonen I, K und P benachteiligt: sie haben, so wie viele andere Testpersonen, nur zwei Zahlenreihen nicht nachsprechen können, bekommen aber im Gegensatz zu den anderen, die noch zwei weitere Zahlenreihen nachsprechen dürfen, *keine weiteren Zahlenreihen* mehr zur Bearbeitung. Abgeschwächt wird dieses Ergebnis wohl durch die Tatsache, dass von diesen Personen nur *eine* Person (D) eine längere Zahlenreihe auch tatsächlich nachsprechen konnte, es also einigermaßen plausibel erscheint, dass auch die hier als „benachteiligt“ bezeichneten Personen dies wahrscheinlich nicht geschafft hätten, dennoch ist es nicht fair, dass ihnen die Chance dazu vorenthalten wird.

Abbildung 6.24 zeigt die Situation nach 12 Items, also nach beiden Aufgaben mit sieben Zahlen:



**Abbildung 6.24:** Abbrüche *Zahlen Nachsprechen vorwärts*: gelöste und nicht gelöste Items **bis zum** Item 12 und **nach dem** Item 12 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 12“ leistungstärksten 17 Testpersonen

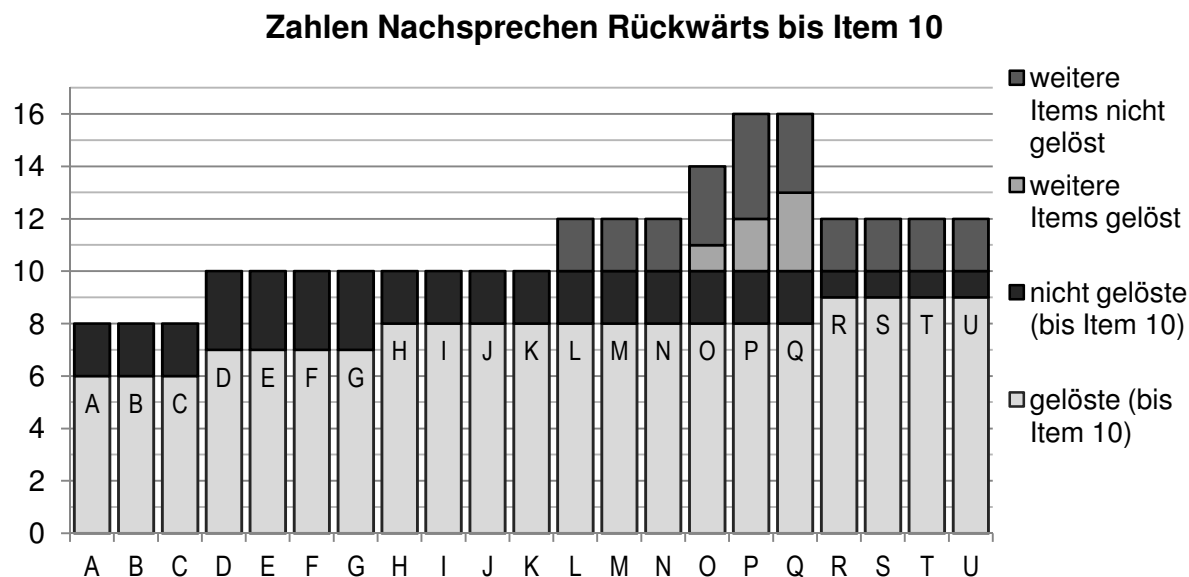


Obwohl die Testpersonen Y und Z genauso wie die Testpersonen AA bis EE nur zwei Zahlenreihen nicht nachsprechen konnten, werden ihnen im Gegensatz zu den anderen keine weiteren Aufgaben vorgegeben! Zu beachten ist, dass die anderen Testpersonen die weiteren Aufgaben auch teilweise lösen konnten! <sup>37</sup>

Ein noch problematischeres Bild zeichnet die Grafik zum *Zahlennachsprechen rückwärts* (Abb. 6.25). In der abgebildeten Situation nach 10 Items, also nach den beiden Zahlenreihen mit fünf Elementen, zeigt sich folgendes Bild:

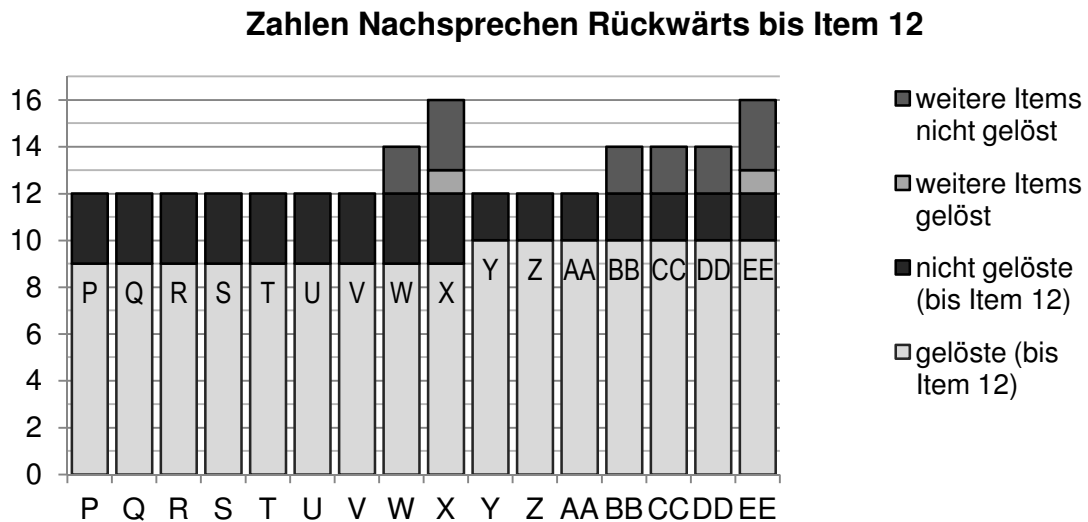
<sup>37</sup> Einer gelösten zusätzlichen Aufgabe entspricht in den allermeisten Altersgruppen und Fähigkeitsbereichen etwa ein Wertpunkt.

**Abbildung 6.25:** Abbrüche Zahlen Nachsprechen rückwärts: gelöste und nicht gelöste Items **bis zum** Item 10 und **nach dem** Item 10 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 10“ leistungsschwächsten 21 Testpersonen



Zehn Testpersonen (H-Q) haben nach zehn Items gleich viele, nämlich acht Punkte: davon dürfen vier Personen (H-K) kein einziges weiteres Item, drei Testpersonen (L-N) zwei weitere Items, eine Testperson (O) vier weitere Items und zwei Testpersonen (P, Q) sogar sechs weitere Items bearbeiten! Ähnliches zeigt sich bei weiteren Personen in Abbildung 6.26:

**Abbildung 6.26:** Abbrüche Zahlen Nachsprechen rückwärts: gelöste und nicht gelöste Items **bis zum** Item 12 und **nach dem** Item 12 vorgegebene gelöste bzw. nicht gelöste Items; die zum Zeitpunkt „Item 12“ leistungsschwächsten 15 und leistungstärksten 11 Testpersonen sind nicht dargestellt



Dargestellt ist die Situation nach 12 Items. Es zeigt sich, dass es auch bei den Personen im mittleren Leistungsfeld bei gleicher Anzahl bisher gelöster Items zu einer unterschiedlichen Anzahl von weiteren Items kommt: Von den Testpersonen, die neun Punkte erlangen konnten, dürfen sieben Testpersonen (P bis V) keine weiteren Items bearbeiten, eine Person (W) noch zwei und eine weitere Person (X) noch vier Items!

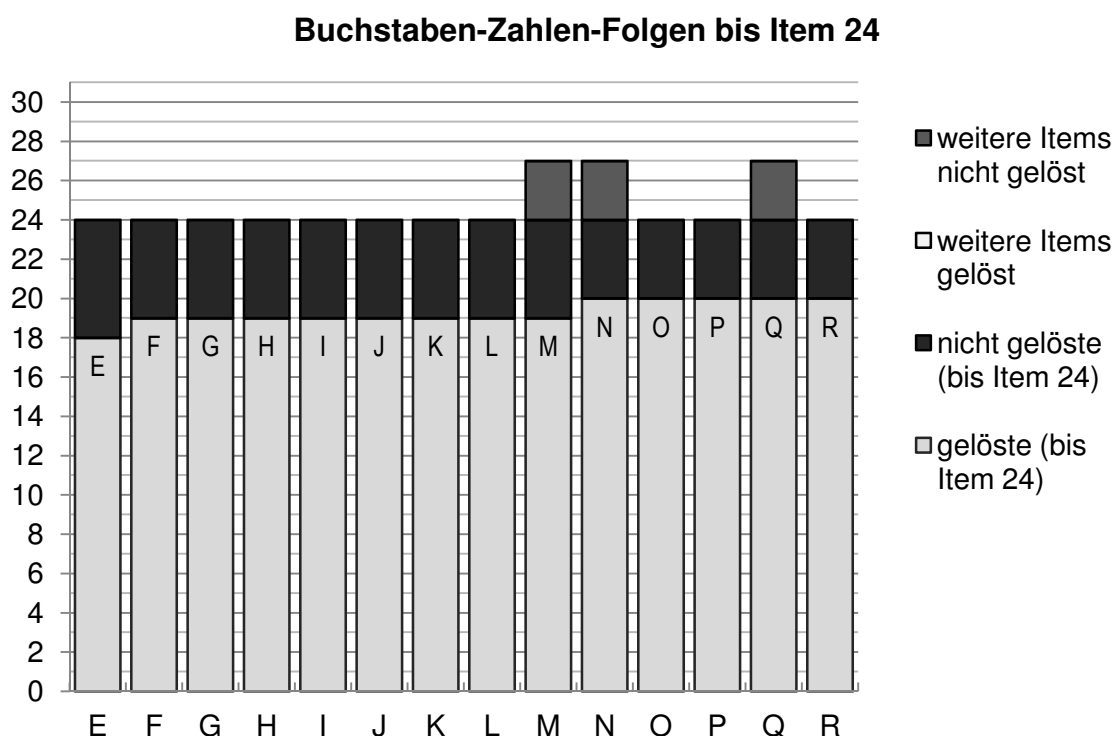
Von den Testpersonen mit 10 Punkten bekommen drei (Y-AA) kein weiteres Item, drei (BB-DD) zwei weitere Items und eine Testperson (EE) vier weitere Items zur Bearbeitung. Diese ungleiche Behandlung ist aber in allen Fällen nicht durch die Zahl der bisher gelösten Items begründet, sondern nur durch die Reihenfolge der Fehler, denn zwei hintereinander nicht gelöste Items, die aber eine *unterschiedliche* Anzahl von Elementen haben (beispielsweise die zweite Zahlenreihe mit fünf Zahlen und die erste Zahlenreihe mit sechs Zahlen), führen zu keinem Abbruch, wohingegen zwei hintereinander nicht gelöste Items, die die selbe Reihenlänge haben, schon zum Abbruch führen!

#### 6.4.9 Abbruchregel Buchstaben-Zahlen-Folgen

Beim Untertest *Buchstaben-Zahlen-Folgen* werden der Testperson Reihen aus Buchstaben und Zahlen vorgesagt, die sie nachsprechen soll, wobei zuerst alle Zahlen der Größe nach geordnet und danach alle Buchstaben in alphabetischer Reihenfolge wiedergegeben werden sollen. Für jede Anzahl von Elementen (von vier bis acht) gibt es

drei Items, die zu einem Aufgabenblock zusammengefasst werden. Für die Reihenlänge „zwei“ gibt es zwei Aufgabenblöcke, also sechs Items, und für die Reihenlänge „drei“ gibt es drei Blöcke, also neun Items. Der Untertest ist abzubrechen, wenn die Testperson alle drei Aufgaben eines Aufgabenblockes nicht fehlerfrei in der richtigen Reihenfolge wiedergeben kann. Dargestellt ist in Abbildung 6.27 die Situation nach 24 vorgegebenen Items, d.h. nach den drei Aufgaben mit jeweils acht Elementen.

**Abbildung 6.27:** Abbrüche *Buchstaben-Zahlen-Folgen*: gelöste und nicht gelöste Items **bis zum** Item 24 und **nach dem** Item 24 vorgegebene gelöste bzw. nicht gelöste Items; die zum Zeitpunkt „Item 24“ leistungsschwächsten 4 und leistungstärksten 23 Testpersonen sind nicht dargestellt

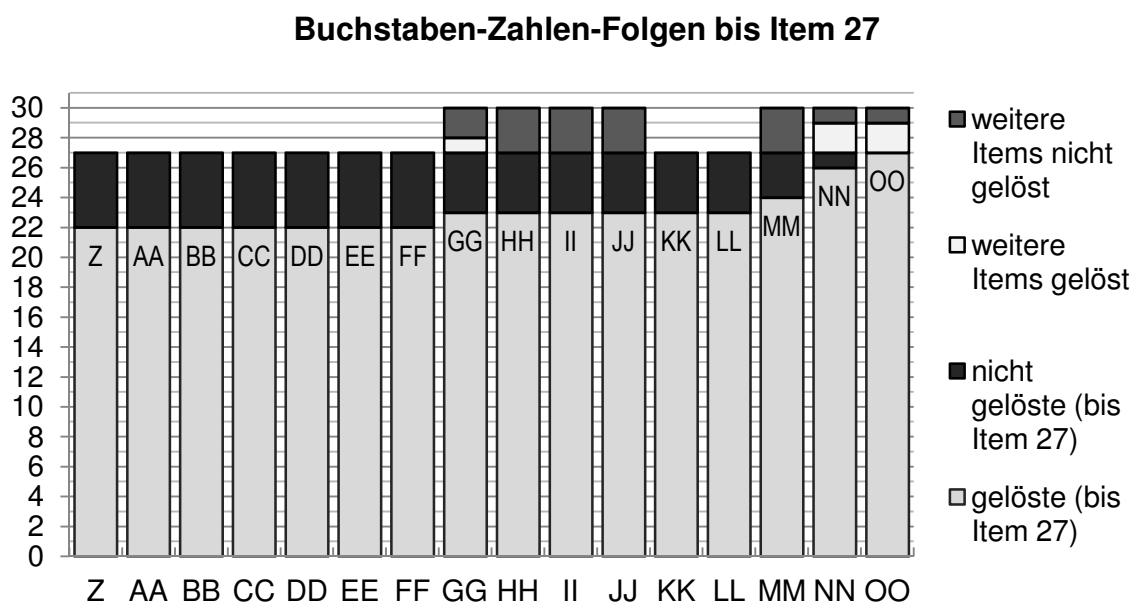


Es ist zu erkennen, dass innerhalb der Blöcke von Testpersonen mit der gleichen Anzahl bisher gelöster Aufgaben einige Testpersonen (M,N,Q) drei Items mehr bearbeiten dürfen als die anderen, diese allerdings nicht lösen können.

Da das Abbruchkriterium hier strenger gewählt wurde, als beim sonst ähnlichen Untertest *Zahlen Nachsprechen*, kommt es auch zu weniger häufigen ungerechten Testabbrüchen wie bei jenem. Dennoch ist anhand von Abbildung 6.28 (nach 27 Items) zu sehen, dass die Testpersonen KK und LL gegenüber den anderen vier Personen mit

22 bisher gelösten Items (GG – JJ) benachteiligt werden, und dass zumindest die Testperson GG eines ihrer zusätzlichen Items auch lösen kann.<sup>38</sup>

**Abbildung 6.28:** Abbrüche *Buchstaben-Zahlen-Folgen*: gelöste und nicht gelöste Items **bis zum** Item 27 und **nach dem** Item 27 vorgegebene gelöste bzw. nicht gelöste Items; die zum Zeitpunkt „Item 27“ leistungsschwächsten 25 Testpersonen sind nicht dargestellt

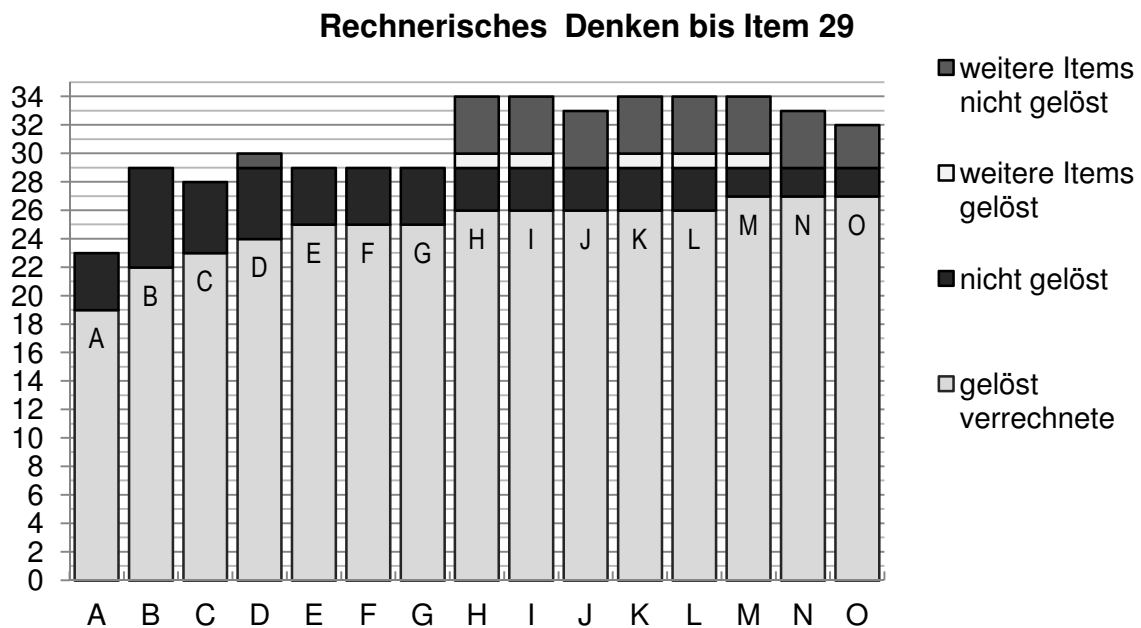


<sup>38</sup> Ein zusätzlich gelöstes Item entspricht in etwa einem Wertpunkt.

#### 6.4.10 Abbruchregel Rechnerisches Denken

Der Untertest *Rechnerisches Denken* ist abubrechen, sobald eine Testperson vier aufeinanderfolgende Items nicht lösen kann.

**Abbildung 6.29:** Abbrüche *Rechnerisches Denken*: gelöste und nicht gelöste Items **bis zum** Item 29 und **nach dem** Item 29 vorgegebene gelöste bzw. nicht gelöste Items; dargestellt sind die zum Zeitpunkt „Item 29“ leistungsschwächsten 15 Testpersonen



In Abb. 6.29 zum Untertest *Rechnerisches Denken* nach 29 Items zeigt sich, dass die Testpersonen mit den Rangplätzen H bis L im Vergleich zu den Testpersonen E bis G, die bisher nur *ein* Item weniger lösen konnten, unverhältnismäßig viel mehr weitere Items zur Bearbeitung bekommen: Sie bekommen nämlich vier oder sogar fünf weitere Items vorgegeben, von denen die meisten Personen je eines lösten, was die Testpersonen E -G zwar nicht stark benachteiligt, aber in geringem Ausmaß durchaus.<sup>39</sup>

### 6.5 Zusammenfassung: Abbruchregeln und Fairness

In diesem Kapitel wurde diskutiert, inwiefern die Abbruchregeln zu Verstößen gegen das Kriterium der Fairness führen. Dieses Problem berührt auch Fragen der Reliabilität, auf die aber in diesem Kapitel nicht eingegangen wurden. Ausgangspunkt der Überlegungen ist, dass eine Abbruchregel nur dann gerechtfertigt ist, wenn die Items

<sup>39</sup> In den meisten Altersgruppen und Fähigkeitsbereichen entsprechen einem zusätzlichen Wertpunkt ein zusätzlich gelöstes Item (selten: zwei Items).

streng monoton nach ihrer Schwierigkeit gereiht sind, also tatsächlich davon auszugehen ist, dass die folgenden Items eine vernachlässigbar geringe Lösungswahrscheinlichkeit für die jeweilige Testperson aufgrund ihrer bisherigen Leistung haben. Außerdem muss gelten, dass die bisherige Leistung adäquat erhoben wird, um abschätzen zu können, ob die Testperson weitere Items lösen können oder nicht. Da im HAWIK-IV die Gesamtzahl der gelösten Items (bzw. dadurch erworbenen Punkte) bei der Berechnung der Fähigkeitsparameter als Maß für die Leistung gilt, ist unverständlich, weshalb dies nicht auch für die Abbruchregel gilt.

Außerdem wurde vermutet, dass es (abgesehen von zufälligen Ereignissen auch) von bestimmten Persönlichkeitseigenschaften abhängt, ob es zu mehreren *aufeinanderfolgend* nicht gelösten Items kommt.

Empirisch konnte gezeigt werden, dass die Items *nicht* streng monoton nach ihrer Schwierigkeit geordnet sind und weiters, dass die Abbrüche teilweise *nicht* mit der bisher gezeigten Leistung (Anzahl der gelösten Items) übereinstimmen, was nahelegt, dass die Abbruchregeln dazu führen, dass einzelne Testpersonen ungerechtfertigter Weise Items nicht bearbeiten dürfen, die sie möglicherweise lösen könnten. Außerdem zeigen sich empirisch Zusammenhänge, die nahelegen, dass bestimmte Persönlichkeitsmerkmale einen Einfluss darauf haben, ob es zu diesen vorzeitigen Abbrüchen kommt, was klar gegen das Kriterium der Fairness spricht.

## 7 DISKUSSION

Diese Arbeit beschäftigt sich mit einzelnen Aspekten der Güte und Verwendbarkeit des HAWIK-IV für die Hochbegabungsdiagnostik im Sinne des Wiener Modells als auch im Sinne einer Klassifikation mittels IQ. Für letztere reichte streng genommen „irgendein“ Verfahren aus, das als Ergebnis einen IQ produziert (Holocher et al. 2008); dennoch kann zumindest im Sinne der Definition von Intelligenz im Sinne Wechslers (siehe Einleitung) gefordert werden, dass eine Intelligenztest-Batterie zum Einsatz kommt, die möglichst breit diejenigen Fähigkeiten prüft, die intelligentes Verhalten zu ermöglichen scheinen, und dabei nicht übermäßig Ressourcen (zeitlich, motivational...) verbraucht. In Hinblick auf die Genauigkeit der Klassifikation ist letztlich vor allem auch die Reliabilität der Intelligenztestbatterie ein wesentliches Kriterium.

Eine Klassifikation in diesem Sinne scheint mit dem HAWIK-IV möglich. Die Untertests prüfen ausreichend breit unterschiedliche kognitive Fähigkeiten, auch wenn die in der Einleitung formulierte Kritik an der Aussagekraft des IQ bestehen bleibt: Der mit dem HAWIK-IV erhobene IQ ist nicht *der* IQ, sondern *ein* IQ, sozusagen der HAWIK-IV-IQ. Ob das bedeutet, dass es demnach auch für jeden unterschiedlichen Intelligenztest eine eigene Hochbegabungsklassifikation nach der Höhe des IQ geben kann und diese unterschiedlichen Klassifikationen beträchtlich auseinanderklaffen, soll hier nicht diskutiert werden (vgl. Schlagheck & Petermann, 2006; Holocher et al, 2008). Die Reliabilität des Gesamt-IQ laut Manual des HAWIK-IV ist mit 0,97 ausreichend, und die zuvor dargestellte Kritik betreffend die empirisch festgestellten niedrigeren Reliabilitätskoeffizienten der Untertests lässt sich nicht ohne weiteres auf den Gesamt-IQ übertragen. Einschränkend zur Eignung des HAWIK-IV zur Hochbegabungs-Klassifikation muss jedoch erwähnt werden, dass der zeitliche Aufwand dieses Tests einzig zur Bestimmung des IQ einigermaßen überhöht erscheint.

Die Hinweise, wonach es zu einer systematischen Benachteiligung von Testpersonen durch die Abbruchkriterien kommt, müssten genauer geprüft werden, da die vorliegenden Daten kein überzeugender Beleg dafür sind. Dass es aber zumindest zu unsystematischen oder zufälligen Benachteiligungen einzelner Testpersonen kommt, konnte in einigen Fällen gezeigt werden.

Für die Verwendung im Rahmen einer Diagnostik nach dem Wiener Modell scheint der HAWIK-IV nicht geeignet zu sein, wie schon Holocher et al. (2008) argumentierten,



wobei sich die hier dargestellten Argumente sich in erster Linie auf die (eingeschränkte) Möglichkeit zur Profilinterpretation beziehen. Zusammenfassend kann gesagt werden: Die drei Indizes *Sprachverständnis*, *Arbeitsgedächtnis* und *Verarbeitungsgeschwindigkeit* sind zumindest inhaltlich eindeutig zu interpretieren, auch wenn der erstgenannte aufgrund der dreikategoriellen Verrechnung mit Persönlichkeitsfaktoren, namentlich „Ausdauer“ konfundiert ist. Der Index *Wahrnehmungsorientiertes logisches Denken* dagegen konnte in der Faktorenanalyse nicht repliziert werden und erscheint damit weder gerechtfertigt noch als inhaltlich eindeutig zu interpretieren. Da insbesondere die dazugehörigen Untertests *Mosaik-Test* und *Bildkonzepte* in der vorliegenden Stichprobe Deckeneffekte und zu geringe Reliabilitätskoeffizienten aufweisen, sind sie für die Profilinterpretation auf Untertestebene nicht geeignet. Ohne auf die weiteren Kritikpunkte (hinsichtlich der Skalierung und der Fairness) einzugehen, lässt sich über die Eignung des HAWIK-IV für eine aussagekräftige Profilinterpretation, wie sie beispielsweise das *Wiener Diagnosemodell zum Hochleistungspotenzial* verlangt, ein negatives Urteil fällen. Die Aussagekraft dieser Bewertung muss allerdings dahingehend eingeschränkt werden, dass die empirischen Ergebnisse zum Großteil auf exploratorischen oder deskriptiven Auswertungen basieren.

Auch dürfen die Ergebnisse nicht auf andere Leistungsbereiche generalisiert werden, da die vorliegende Stichprobe sich aus Kindern und Jugendlichen mit fraglicher Hochbegabung zusammensetzte, die sowohl hinsichtlich ihrer kognitiven Fähigkeiten überdurchschnittlich waren, als auch dadurch selektiert waren, dass sie alle zuvor in der Test- und Beratungsstelle Beratung gesucht hatten. Inwiefern das Auswirkungen auf die Testergebnisse hatte, ist unbekannt.

Dies leitet zu der Kritik an dieser Arbeit über:

Die vorliegende Arbeit und ihre Fragestellungen entstanden *nach* der Datenerhebung und mit der Erfahrung *aus* der Datenerhebung „im Hinterkopf“, weswegen alle darin enthaltenen inferenzstatistischen Aussage nicht den Status einer Hypothesenprüfung in Anspruch nehmen können. Darüber hinaus ist der Auswahl der behandelten Gütekriterien eine gewisse Beliebigkeit zu unterstellen, dahingehend, dass genau *die* Kritik dargestellt wurde, die entweder schon im Zuge der Testungen auffiel, oder sich eben in den Daten „finden ließ“. Dieses Vorgehen erscheint dem Verfasser (im Sinne einer Exploration) verantwortbar, zumal mehrfach explizit darauf Bezug genommen

wurde. Dennoch schränkt es die Aussagekraft der hier dargestellten Ergebnisse drastisch ein.

Ebenso eingeschränkt werden die Ergebnisse durch die geringe Stichprobengröße: viele der dargestellten statistischen Ergebnisse sind mit einer relativ großen Irrtumswahrscheinlichkeit behaftet, bzw. sind sie in den seltenen Fällen, in denen ein  $\alpha$ -Niveau festgesetzt wurde, statistisch nicht signifikant, obgleich das Ergebnis – deskriptiv betrachtet – bedeutsam erscheint.

Im Zuge der Auswertungen und Überlegungen wurden Hypothesen gebildet, die eine genauere Datenerhebung notwendig gemacht hätten: exemplarisch sei hier angeführt, dass es sinnvoll gewesen wäre, auch nach den Abbrüchen (lt. Abbruchkriterien) noch weitere Items vorzugeben, um nicht nur die Voraussetzungen der Abbruchkriterien zu überprüfen, sondern auch zu überprüfen, ob *tatsächlich* Items vorenthalten wurden, die andernfalls gelöst worden wären.

Zukünftige Untersuchungen könnten ausgehend von den hier dargestellten Befunden hypothesengeleitet einzelne Kritikpunkte genauer beleuchten. Neben der im letzten Absatz beschriebenen Vorgangsweise hinsichtlich der Abbruchkriterien wären vor allem Studien sinnvoll, in denen die Konfundierung der dreikategoriell zu verrechnenden Untertests mit Persönlichkeitsfaktoren in Gegenüberstellung zu den (nachträglich) dichotomisierten Skalen genauer untersucht werden kann – gegebenenfalls sogar mit in Rahmen von Experimenten kontrolliert veränderten Motivationsbedingungen und Persönlichkeitszuständen.

Als weiterer Kritikpunkt kann und soll die tendenziell kritische Perspektive dieser Arbeit angemerkt werden: eine umfassende Würdigung des HAWIK-IV (hinsichtlich aller positiv zu bewertenden Charakteristika) wurde vom Verfasser nicht intendiert und ist an anderer Stelle bereits vorgenommen worden (z.B. in Petermann & Petermann, 2007, Daseking et al., 2007).

## 8 ZUSAMMENFASSUNG

Die vorliegende Arbeit beschäftigt sich kritisch mit der Testkonstruktion und Eignung des HAWIK-IV für die Hochbegabungsdiagnostik, sowohl im Sinne des Wiener Modells als auch im Sinne einer Klassifikation mittels IQ. Insbesondere wurde der HAWIK-IV hinsichtlich der Gütekriterien Validität, Reliabilität und Fairness untersucht. Zu diesem Zweck wurden sowohl Items inhaltlich analysiert als auch die Daten einer 41 Personen umfassenden Stichprobe exploratorisch ausgewertet.

Die Ergebnisse betreffend Validität lassen sich so zusammenfassen: aufgrund inhaltlicher Überlegungen und faktorenanalytischer Ergebnisse erscheint die Zuordnung des Untertests *Rechnerisches Denken* zum Index *Arbeitsgedächtnis* problematisch, da er (neben der Merkfähigkeit) offensichtlich auch Rechenfähigkeit prüft, was sich bei inhaltlicher Betrachtung der Items und anhand faktorenanalytischer Ergebnisse zeigt; auch schien der Verzicht auf die explizite Prüfung numerischer Intelligenz im Rahmen einer Intelligenztestbatterie für Kindern und Jugendliche kaum nachvollziehbar.

Die von Petermann und Petermann (2007) publizierte Faktorenstruktur, aus der die Bestimmung der Indexwerte aus den Untertestergebnissen resultiert, konnte nur teilweise repliziert werden. Insbesondere der Index *Wahrnehmungsgebundenes logisches Denken* konnte in der Faktorenanalyse der vorliegenden Daten nicht wiedergefunden werden; auch inhaltlich weist dieser Index und seine Untertests (teilweise aufgrund von Konfundierungen mit Speedkomponenten) eine mangelhafte inhaltliche Interpretierbarkeit auf. Die postulierte Prüfung fluider Intelligenz durch diese Untertests konnte aufgrund mangelnden statistischer Zusammenhänge zum Intelligenztest CFT-20 R nicht bestätigt werden.

In Hinblick auf mögliche Konfundierungen mit Persönlichkeitsvariablen wurde auch die dreikategorielle Verrechnung der Kerntests des Index *Sprachverständnis* untersucht. Diese Konfundierung konnte inhaltlich plausibel gemacht und anhand der vorliegenden Daten zumindest für die Persönlichkeitsvariable *Ausdauer in der Testsituation* dargestellt werden.

Im Kapitel zur Reliabilität wurde gezeigt, dass die Berechnung der Reliabilität nach der Split-half-Methode empfindlich gegenüber Verzerrungen (nach oben) ist, wenn ein Untertests eine relativ große Anzahl sehr leichter oder sehr schwieriger Items aufweist,

wie es bei einigen Untertests des HAWIK-IV der Fall ist. Aufgrund der dargestellten Überlegungen und entsprechender Datenanalyse konnten berechtigte Zweifel daran formuliert werden, dass die bei Petermann und Petermann (2007) publizierten Split-half-Reliabilitäten zuverlässige Schätzungen der Messgenauigkeit einiger Untertests darstellen. Außerdem konnten z. T. starke Deckeneffekte aufgezeigt werden. Da darüber hinaus die Messgenauigkeit einiger anderer Untertests schon anhand der bei Petermann und Petermann (2007) publizierten Reliabilitätsangaben als zu gering für die Interpretation als Einzelskala angesehen werden können, erscheint der HAWIK-IV für eine Profilinterpretation nur wenig geeignet.

Weiters wurden Kritikpunkte an der Praxis der Abbruchkriterien formuliert. So wurde gezeigt, dass die logische Voraussetzung dieser Abbruchkriterien, nämlich stetig ansteigende Itemschwierigkeiten, bei einigen Untertests keinen Niederschlag in den empirischen Lösungshäufigkeiten fanden, woraus sich berechtigte Zweifel an der Erfüllung eben dieser Voraussetzung ergeben. Darüber hinaus wurde dargestellt, dass die Abbruchkriterien in vielen Fällen im Widerspruch zu den Verrechnungsvorschriften zur Bestimmung der Normwerte liegen, was ungerechtfertigte Abbrüche wahrscheinlich macht. Anhand der Daten konnte auch belegt werden, dass die Frühzeitigkeit der Abbrüche in statistischem Zusammenhang zu einigen Persönlichkeitsmerkmalen steht (beispielsweise *Anstrengungsbereitschaft* sowie *Ausdauer in der Schule* und *Schulischer Ehrgeiz*). Davon ausgehend wurde argumentiert, dass es aufgrund der Abbruchkriterien zu Verstößen gegen das Kriterium der Fairness kommt.

Obwohl die Ergebnisse nicht im Sinne einer Hypothesenprüfung verstanden werden können, liefern sie genug Anhaltspunkte dafür, den HAWIK-IV vorläufig als nicht geeignet für eine Diagnostik im Sinne der *Wiener Diagnosemodells zum Hochbegabungspotenzial* zu bewerten, für die (nicht empfohlene!) Klassifikation nach einem festgesetzten IQ-Cut-Off-Wert jedoch als geeignet.

## 9 LITERATUR

- Amelang, M., Bartussek, D., Stemmler, G., & Hagemann, D. (2006). *Differentielle Psychologie und Persönlichkeitsforschung* (6., vollständig überarbeitete Auflage). Stuttgart: Kohlhammer
- Bortz, J. & Lienert, G.A. (2003). *Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben*. Berlin: Springer
- Cattell, R.B., Weiß, R.H. & Osterland, J. (1997) *Grundintelligenztest Skala 1*. Göttingen: Hogrefe
- Daseking, M., Petermann, U. & Petermann, F. (2007). Intelligenzdiagnostik mit dem HAWIK-IV. *Kindheit und Entwicklung*, 16, 250-259
- Dilling, H., Mombour, W. & Schmidt, M.H. (2000). *Internationale Klassifikation psychischer Störungen: ICD-10, Kapitel V (F); klinisch-diagnostische Leitlinien*. Bern: Huber
- Fisseni, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik mit Hinweisen zur Intervention*. Göttingen: Hogrefe
- Grosjean, S. (2003). *Liebeserklärung an die Psychologie*. Unveröffentlicht. Verfügbar unter (30.1.2009): <http://www.psych.stgr.ch/Liebeserklaerung%20an%20die%20Psychologie.pdf>
- Holling, H. & Kanning, U.P. (1999). *Hochbegabung. Forschungsergebnisse und Fördermöglichkeiten*. Göttingen: Hogrefe
- Holocher-Ertl, S., Kubinger, K.D. & Hohensinn, C. (2006). Zur Definition von Hochbegabung ist die Höhe des IQ zwar Konvention aber völlig ungeeignet: Ein neues Diagnosemodell im Spannungsfeld von Hochbegabung und Hochleistung. In B. Gula, R. Alexandrowicz, S. Strauß, E. Brunner, B. Jenull-Schiefer & O. Vitouch (Hrsg.), *Perspektiven psychologischer Forschung in Österreich. Proceedings zur 7. Wissenschaftlichen Tagung der Österreichischen Gesellschaft für Psychologie* (S. 444-451). Lengerich: Pabst.
- Holocher-Ertl, S., Kubinger, K.D., Hohensinn, C. (2008). Hochbegabungsdiagnostik: HAWIK-IV oder AID 2. *Kindheit und Entwicklung*, 17, 99-106.

Kubinger, K.D. (Hrsg.) (1983). *Der HAWIK. Möglichkeiten und Grenzen seiner Anwendung*. Weinheim: Beltz

Kubinger, K.D. (2006). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe

Kubinger, K.D. & Wurst, E. (2000). *Adaptives Intelligenz Diagnostikum – Version 2.1*. Göttingen: Beltz

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Psychologie-Verlags-Union

Petermann, F. & Petermann, U. (2007). *HAWIK®-IV: Hamburg-Wechsler-Intelligenztest für Kinder-IV. Übersetzung und Adaption der WISC-IV® von David Wechsler*. Göttingen: Hogrefe

Preusche, I. (2007). *Fairness von Intelligenztests bei Kindern und Jugendlichen*. Saarbrücken: VDM Verlag Dr. Müller

Rasch, D. & Kubinger, K.D. (2006). *Statistik für das Psychologiestudium – Mit Softwareunterstützung zur Planung und Auswertung von Untersuchungen sowie zu sequentiellen Verfahren*. Heidelberg: Spectrum.

Schlagheck, W. & Petermann, F. (2006). Hochbegabungsdiagnostik mit dem HAWIK-III und AID 2. *Kindheit und Entwicklung*, 15, 93-99

Schubhart, S. (2008). *Katamnestic Validierung des "Wiener Diagnosemodells zum Hochleistungspotenzial"*. Unveröffentlichte Diplomarbeit, Universität Wien

Seitz, W. & Rausche, A. (2004) *Persönlichkeitsfragebogen für Kinder zwischen 9 und 14 Jahren. 4. überarbeitete und neu normierte Auflage*. Göttingen: Hogrefe

Steurer, O. (1988). *HAWIK und HAWIK-R: testtheoretische Analysen des HAWIK und seiner revidierten Form als Wiederholungsstudie und Weiterführung der Arbeit von Kubinger (1983): "Der HAWIK - Möglichkeiten und Grenzen seiner Anwendung"*. Unveröffentlichte Dissertation, Universität Wien

The free dictionary: <http://de.thefreedictionary.com/Fairness>, (20.10.08)

Weiß R.H., (2006). *Grundintelligenztest Skala 2 – Revision*. Göttingen: Hogrefe

Wikipedia: <http://de.wikipedia.org/wiki/Fairness>, (20.10.08)

*Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit einzuholen. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.*

## 10 ANHANG

**Tabelle A.1:** Trennschärfen der Items der dreikategoriellen und der dichotomisierten („schwierig“) Skala des Untertests *Wortschatz-Test* (Die Items 1 bis 4 wurden keiner Tpn vorgegeben und werden daher nicht angeführt)

Item-Nr.	Anzahl der Tpn	Itemschwierigkeit		Trennschärfe nach Pearson		Trennschärfe nach Spearman	
		Drei-kategoriell	Dichotom_schwierig	Drei-kategoriell	Dichotom_schwierig	Drei-kategoriell	Dichotom_schwierig
WT5	5	0,90	0,8	0,63	0,70	0,71	0,71
WT6	5	1,00	1				
WT7	27	1,00	1				
WT8	27	1,00	1				
WT9	41	0,99	0,98	0,31	0,31	0,26	0,27
WT10	41	0,99	0,98				
WT11	41	1,00	1				
WT12	41	0,98	0,98	0,26	0,21		0,19
WT13	41	0,93	0,88	0,26	0,22		0,21
WT14	41	0,93	0,85				
WT15	41	0,93	0,85	0,36	0,41	0,32	0,37
WT16	41	0,94	0,9	0,31	0,28	0,34	0,29
WT17	41	0,98	0,95	0,36	0,36	0,32	0,32
WT18	41	0,95	0,95	0,26	0,24		0,24
WT19	41	0,83	0,83	0,66	0,60	0,59	0,56
WT20	41	0,92	0,9	0,30	0,34	0,32	0,33
WT21	41	0,78	0,68				
WT22	41	0,84	0,8	0,51	0,50	0,52	0,49
WT23	41	0,61	0,46	0,46	0,36	0,45	0,35
WT24	41	0,55	0,51	0,75	0,78	0,76	0,79
WT25	41	0,65	0,56	0,62	0,45	0,56	0,42
WT26	41	0,70	0,68	0,65	0,61	0,59	0,57
WT27	41	0,84	0,73	0,39	0,34	0,40	0,36
WT28	40	0,66	0,58	0,64	0,49	0,57	0,45
WT29	40	0,69	0,53	0,59	0,36	0,48	0,33
WT30	40	0,56	0,5	0,64	0,61	0,67	0,64
WT31	40	0,85	0,83				
WT32	39	0,41	0,18	0,45	0,49	0,51	0,51
WT33	39	0,71	0,59	0,57	0,42	0,43	0,36
WT34	39	0,46	0,44	0,37	0,33	0,34	0,32
WT35	39	0,17	0,15	0,41	0,43	0,47	0,46
WT36	39	0,60	0,44	0,58	0,49	0,65	0,51



**Tabelle A.2:** Trennschärfen der Items der dreikategoriellen und der dichotomisierten („schwierig“) Skala des Untertests *Allgemeines Verständnis* (Die Items 1 und 2 wurden keiner Tpn vorgegeben und werden daher nicht angeführt)

Item-Nr.	Anzahl der Tpn	Itemschwierigkeit		Trennschärfe nach Pearson		Trennschärfe nach Spearman	
		Drei-kategoriell	Dichotom-schwierig	Drei-kategoriell	Dichotom-schwierig	Drei-kategoriell	Dichotom-schwierig
AV3	0	0,98			0,36	0,27	0,31
AV4	0	0,92		0,29	0,50	0,34	0,44
AV5	5	0,99	0,8		0,29		0,26
AV6	5	1	1				
AV7	27	0,91	1	0,32	0,41	0,37	0,38
AV8	27	0,82	1	0,54	0,38	0,45	0,40
AV9	41	0,79	0,98	0,33	0,36	0,27	0,36
AV10	41	0,8	0,98	0,27	0,38	0,31	0,40
AV11	41	0,61	1	0,46	0,49	0,54	0,52
AV12	41	0,84	0,98		0,42		0,40
AV13	41	0,35	0,88	0,58	0,66	0,63	0,67
AV14	41	0,5	0,85	0,61	0,64	0,73	0,65
AV15	41	0,66	0,85	0,67	0,69	0,67	0,72
AV16	41	0,53	0,9	0,42	0,16	0,40	0,17
AV17	41	0,67	0,95	0,59	0,56	0,58	0,56
AV18	41	0,28	0,95	0,57	0,45	0,64	0,44
AV19	41	0,31	0,83	0,59	0,30	0,69	0,27
AV20	41	0,32	0,9	0,65	0,32	0,65	0,29
AV21	41	0,27	0,68	0,63	0,42	0,71	0,48



## LEBENS LAUF

seit 2008	Praktikum und weitere Mitarbeit an der <i>Test- und Beratungsstelle des Arbeitsbereichs Psychologische Diagnostik</i> , Fakultät für Psychologie, Universität Wien: psychologische Diagnostik bei Kindern und Jugendlichen, Betreuung des AID 2 – Zertifizierungskurses, Projektmitarbeit „Intelligenztestungen bei Kindern und Jugendlichen mit türkischer Muttersprache“
seit 2003	Diplomstudium Psychologie an der Universität Wien
2001 - 2003	Behindertenpädagoge in der Wohngemeinschaft Redtenbachergasse des Vereins <i>Jugend am Werk</i> , Wien
1998 - 2001	Behindertenpädagoge und Leitungsstellvertreter in der Wohngemeinschaft Senefeldergasse des <i>Verein TRIAS</i> , Wien
1995 - 1998	Ausbildung zum „diplomierten Behindertenpädagogen / diplomierten heilpädagogischen Fachbetreuer“ in der <i>Lehranstalt für heilpädagogische Berufe</i> , Wien 2
1997 - 1998	Behindertenbetreuer in einer teilbetreuten Wohngemeinschaft der <i>Lebenshilfe Niederösterreich</i> , Wiener Neustadt
1995 - 1996	Behindertenbetreuer im „Haus Franciscus“ der <i>Caritas Wien</i>
1993 - 1995	Erzieher im <i>Kinderheim der Caritas</i> „Am Himmel“, Wien, teilweise im Rahmen des ordentlichen Zivildienstes
1985 - 1993	Bundesgymnasium Wien IX, Wasagasse
25.11.1974	geboren in Wien